Summer 2007

Designing multimodal interaction for the visually impaired

Xiaoyu Chen
New Jersey Institute of Technology

ABSTRACT

DESIGNING MULTIMODAL INTERACTION FOR THE VISUALLY IMPAIRED

by
Xiaoyu Chen

Although multimodal computer input is believed to have advantages over unimodal input, little has been done to understand how to design a multimodal input mechanism to facilitate visually impaired users' information access.

This research investigates sighted and visually impaired users' multimodal interaction choices when given an interaction grammar that supports speech and touch input modalities. It investigates whether task type, working memory load, or prevalence of errors in a given modality impact a user's choice. Theories in human memory and attention are used to explain the users' speech and touch input coordination.

Among the abundant findings from this research, the following are the most important in guiding system design: (1) Multimodal input is likely to be used when it is available. (2) Users select input modalities based on the type of task undertaken. Users prefer touch input for navigation operations, but speech input for non-navigation operations. (3) When errors occur, users prefer to stay in the failing modality, instead of switching to another modality for error correction. (4) Despite the common multimodal usage patterns, there is still a high degree of individual differences in modality choices.

Additional findings include: (1) Modality switching becomes more prevalent when lower working memory and attentional resources are required for the performance of other concurrent tasks. (2) Higher error rates increases modality switching but only

under duress. (3) Training order affects modality usage. Teaching a modality first versus second increases the use of this modality in users' task performance.

In addition to discovering multimodal interaction patterns above, this research contributes to the field of human computer interaction design by: (1) presenting a design of an eyes-free multimodal information browser, (2) presenting a Wizard of Oz method for working with visually impaired users in order to observe their multimodal interaction.

The overall contribution of this work is that of one of the early investigations into how speech and touch might be combined into a non-visual multimodal system that can effectively be used for eyes-free tasks.

DESIGNING MULTIMODAL INTERACTION FOR THE VISUALLY IMPAIRED

by
Xiaoyu Chen

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems

Department of Information Systems

August 2007

APPROVAL PAGE

DESIGNING MULTIMODAL INTERACTION FOR THE VISUALLY IMPAIRED

Xiaoyu Chen

July 16, 2007

Dr. Marilyn M. Tremaine, Dissertation Advisor — Date
Professor Emerita of Information Systems, NJIT

7/16/2007

Dr. Murray Turoff, Committee Member — Date
Distinguished Professor Emeritus of Information Systems, NJIT

7/16/2007

Dr. Quentin Jones, Committee Member — Date
Assistant Professor of Information Systems, NJIT

7/30/2007

Dr. Brian Whitworth, Committee Member — Date
Senior Lecturer of Information Systems, Massey University, New Zealand

8/8/07

Dr. Ephraim Glinert, Committee Member — Date
Program Director of Human-Computer Interaction and Universal Access, National Science Foundation
Professor Emeritus of Computer Science, Rensselaer Polytechnic Institute

BIOGRAPHICAL SKETCH

Author: Xiaoyu Chen

Degree: Doctor of Philosophy

Date: August 2007

Undergraduate and Graduate Education:

- Doctor of Philosophy in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2007

- Master of Science in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2006

- Bachelor of Engineering in Management Information Systems,
  Beijing University of Aeronautics and Astronautics, Beijing, P. R. China, 1998

Major: Information Systems

Presentations and Publications:

Chen, X., Tremaine, M. M., Lutz, R., Chung, J. W., & Lacsina, P. (2006). AudioBrowser: A mobile browsable information access for the visually impaired. *International Journal of Universal Access in the Information Society*, 5(1), 4-22.

Chen, X. and Tremaine, M. M. (2006). Patterns of multimodal input usage in non-visual information navigation. In *Proceedings of the thirty ninth Hawaii international conference on system sciences (HICSS)*, Kauai, Hawaii, January 2006.

Chen, X. and Tremaine, M. M. (2005). User error handling strategies on a non-visual multimodal interface: preliminary results from an exploratory study. In *Proceedings of the eleventh American conference on information systems (AMCIS)*, Omaha, Nebraska, August 2005.

Chen, X. (2005). Designing non-visual multimodal dialogues to support information access for the visually impaired. In *Extended abstract of proceedings of the eleventh American conference on information systems (AMCIS)*, Omaha, Nebraska, August 2005.

Chen, X. and Tremaine, M. M. (2005). Multimodal user input patterns in a non-visual context. In *Proceedings of the seventh international ACM SIGACCESS conference on computers and accessibility (ASSETS)*, Baltimore, Maryland, October 2005.

Chen, X. (2005). Mixed-mode dialogue information access for the visually impaired. *ACM SIGACCESS Accessibility and Computing*, 81, 16-19.

Chen, X., Chung, J. W., Lacsina, P., & Tremaine, M. M. (2004). Mobile browsable information access for the visually impaired. In *Proceedings of the tenth American conference on information systems (AMCIS)*, New York, New York, August 2004.

Chen, X. and Wang, Y. (2004). Using synchronous chat to improve online learning experience. In *Proceedings of the tenth American conference on information systems (AMCIS)*, New York, New York, August 2004.

Chen, X., Lacsina, P., & Tremaine, M. M. (2003). Designing non-visual bookmarks for mobile PDA users. In *Proceedings of the ninth American conference on information systems (AMCIS)*, Tampa, Florida, August 2003.

To my beloved parents

献给从未间断过支持我，爱我的父亲和母亲

ACKNOWLEDGMENT

I would like to express my deepest appreciation to Dr. Marilyn Mantei Tremaine, who not only served as my research supervisor, providing valuable and countless resources, insight, and intuition, but also constantly gave me support, encouragement, and reassurance. Her technical and editorial advice was essential to the completion of this dissertation. Special thanks are given to Dr. Brian Whitworth, Dr. Ephraim Glinert, Dr. Quentin Jones and Dr. Murray Turoff for actively participating in my committee and providing many precious comments and suggestions for this work.

Special thanks are given to the New Jersey Commission for the Blind and Visually Impaired and Joseph Ruffalo at National Federation of the Blind, New Jersey, for their strong support in finding participants for this research. My deepest appreciation is given to the twenty visually impaired participants who not only spent many hours participating and providing invaluable opinions, but also gave me tremendous encouragement. Their unconditional and warm-hearted support was essential to the success of this work.

I am also grateful to Lynn Cherny, Teresa Bleser and Jason Winstanley, my former manager and current managers at Autodesk Inc. It was their generous support and understanding that allowed me to finish my doctoral research while working full-time.

I would also like to give my deepest appreciation to John Visicaro, my research team colleague who helped me in finding participants and preparing for the research experiment, and encouraged me during the tough period, and Janice Ortiz, who volunteered to be my system design consultant and provided precious insights and

suggestions. John and Janice passed away before this dissertation was finished. Their friendship and generosity will live in my heart forever.

My thanks also go to Jaewoo Chung, who provided numerous inspiring design ideas and helped in system implementation during an early stage of the research.

I would also like to thank my friends, Xiang, Mengmeng, Xuezeng, Peng, Jing, Aohua and Weiwei for their friendship, encouragement and experience sharing during the highs and lows on the road pursuing my doctoral degree.

My special appreciation is given to Patrick Lacsina, who not only helped me in the implementation of the experiment system but also proofread many chapters of this dissertation. Moreover, his tremendous and continuous support during the pursuit of my doctoral degree is fundamental and essential.

Last, but not least, my deepest appreciation is given to my parents and brother, who gave me their love, encouragement, belief and support unconditionally. Without them standing behind me, I would not have walked through the long and tough way.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

Researchers' early efforts to provide information access for visually impaired individuals through the use of computing technologies can be traced back to the 1980s' when the first screen readers were generated to work with DOS and UNIX operating systems (Thatcher, 1994). Researchers' continuous efforts in this field in the past two decades have resulted in systems that read stored documents, operating system interfaces, and Web pages; systems that provide access to special information types such as graphs, tables, and mathematical notations; systems that travel with the user to provide portable information access; and innovations in input and output technologies and techniques customized for the visually impaired.

Although these systems have significantly increased information and computer systems accessibility, limitations exist in their input designs. For example, keyboards and keypads are broadly used to provide a large amount of functions through single or combined keystrokes (e.g., Asakawa and Itoh, 1998; Braille n Speak by Freedom Scientific). Issues related to these designs are the high memorization load required to efficiently use these mechanisms and the requirement that users sequentially access information using arrow keys to listen to and step through the information structure until the desired information is found. A touch screen using a "touch-to-hear" mechanism can break this sequence of information presentation (Roth et al. 2000), but the cost of the device is high. A set of head gestures can be easily learned and used to also skip through information (Brewster et al. 2003). However, using such gestures in a public environment may make the user look and feel awkward.

In general, different input modalities provide different advantages to users but also introduce serious limitations. at the same time. One solution to this problem is to design an integrated multimodal input system for the visually impaired that appropriately matches the input mechanism to the user task.

Well-designed multimodal input mechanisms have a great potential to improve information and systems accessibility. By complementing each other, multiple input modalities can yield a "highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other" (page 576, Oviatt, 1999b). Having multiple input modalities can create a failsafe system in which one modality can be used to correct errors that occur in the other input modality (Oviatt, 1996). Moreover, multimodal interfaces are expected to support the natural coordination of speech and hand motions because, as linguists have uncovered, there is a close synchrony between speech and hand gestures in human communication. In fact, speech and hand motions have been found to be inseparable units expressing different aspects of the same conceptual content in the communication (McNeil, 2000).

Coordinated speech and hand inputs can provide different advantages for the visually impaired. Speech input can provide a comprehensive grammar without limiting the user to learning physical devices and complex motor actions. Speech input can provide fast and direct access to system functions without the need for users to browse menus. Hand input allows visually impaired users to take advantage of their sense of touch acuity. Hand operations also do not interrupt a user's comprehension of computer speech output. Research on multimodal input on GUIs indicated that combined gesture and speech inputs improved a system's speech understanding by using gesture to

disambiguate what was meant by speech (Bolt, 1987; Hauptmann and MacAvinney, 1993). Combined gesture and speech also sped up the interaction process since many gestures could be carried out simultaneously with speech (Thorisson et al. 1992).

In addition to the above research, theories of cognitive psychology also imply a promising future for multimodal interaction for visually impaired users. Theories in human attention specify that there exist multiple pools of attentional resources, each of which processes specific types of information. Multiple tasks can be performed at the same time as long as they require separate pools of resources (Wickens, 1980 and 1984). There is also research showing that verbal information and spatial movement are processed by different pools of resources (Wickens and Liu, 1988). The theory in human working memory suggests that working memory consists of three components, two of which process verbal and spatial information separately, and the third which integrates the processed verbal and spatial information (Baddeley and Hitch, 1974).

However, little in practice has been done on designing multimodal speech and hand input dialogues for visually impaired users. This is because much prior work on multimodal inputs assumes that hand operations in multimodal input are performed using hand-eye coordination, a skill not available to the visually impaired. Furthermore, prior work on dialogues of integrated speech and hand input mainly focused on two application domains: interactive map based tasks (e.g., Oviatt, 1997; Oviatt et al. 2003), and cross-modal error correction in speech recognition systems (e.g., Sears et al. 2003; Suhm et al. 2001). Empirical research is needed to help us understand how to unleash the power of multimodal input to facilitate visually impaired users' non-visual interaction.

The goal of this research is to provide such empirical results that lead to the formation of design principles. The research was conducted on AudioBrowser (Chen et al. 2006), a non-visual information access with parallel speech and touch user input. This thesis reports users' non-visual multimodal interaction patterns captured through two studies using AudioBrowser: an exploratory study with sighted users aiming at understanding the design problems and drafting the scope of the research, and a controlled experiment with visually impaired users evaluating and extending the understandings obtained from the exploratory study. It discusses discovered interaction behaviors basing them on cognitive psychology theories. This work also reports a research method that uses Wizard of Oz simulation to capture visually impaired users' interaction with a speech recognition system.

The rest of this thesis is organized into the following general sections: (1) a review of related research, (2) the research questions addressed by the thesis and a description of the AudioBrowser system, (3) the design, results and discussion of the exploratory study with sighted users, (4) the design, results and discussion of the controlled experiment with visually impaired users, and (5) the contributions.

CHAPTER 2

RELATED WORK

2.1 Overview

This literature review contains three general parts.

The first part reviews existing designs of information access for visually impaired users. It consists of the following subsections: (1) design of screen readers, (2) non-visual tools for navigation in hyper spaces, (3) designs that read specific information types, and (4) enriched audio outputs used for information representation. The purposes of this part are to obtain an overview of the available designs for the targeted users and advantages and disadvantages of those designs, and to identify the requirements and design issues that have not been addressed in current designs. The review results in the suggestion of a new form of interface dialogue that involves multimodal speech and hand input.

The second part reviews work done on multimodal input dialogue designs on graphical user interfaces, since multimodal input on non-visual interfaces are not yet available. This part focuses especially on interfaces containing speech and hand gesture inputs. It consists of the following subsections: (1) lessons learned from gesture input design, (2) lessons learned from designs of speech dialogues, (3) the advantages of combined speech and gesture inputs, and (4) findings and interpretations from existing studies on multimodal input on graphical user interfaces.

The third part reviews the theories and empirical studies in cognitive psychology relevant to multimodal interaction design. It has two subsections: (1) human attention, especially models if attention and theories about attention allocation to information

conveyed through different modalities concurrently, and (2) working memory, especially how multimodal information is processed in the working memory.

A summary of the literature review is provided at the end of this section.

2.2 Designs for the Visually Impaired

2.2.1 Design and History of Screen Readers

Software integrated with a speech synthesizer to read content objects aloud and inform users about events on computer screens has been given a general name: screen readers. Older versions of screen readers worked on text-based operating systems such as DOS and UNIX. Newer versions work on Graphic User Interfaces (GUIs), which can specify icons, menus, control buttons, events of dialogue boxes, and so on. Screen readers provide access to stored documents, web contents, and special information styles (e.g., software development environments, tables and mathematical notations). A selected collection of screen readers is described in this section.

In the late 1970's, a prototype called SAID (Synthetic Audio Interface Driver) was created by IBM (Thatcher, 1994). This was the first attempt by computer scientists to make electronic information accessible to blind users. The main idea was to transform text displayed through the IBM 3377 terminals to auditory output using synthesized speech. The device was a modified terminal including a twelve-key keypad and a multi-lingual voice system. Users had limited control using the keypad and the required hardware was very costly. The prototype was not turned into a product because of the unavailability of many needed technologies.

Following this attempt, in 1988, when related technologies were significantly improved, IBM developed Screen Reader/DOS that transformed the visual display of DOS to speech output (Thatcher, 1994). It included an 18-key keypad and a Profile Access Language (PAL) that could make changes in the keypad functions and the speech output style.

In 1994, when GUIs started to be widely adopted, IBM created Screen Reader/2 for the windows applications running on its OS/2 operating system (Thatcher, 1994). The goal was to convert a large collection of functions on the GUI to speech output. Using an eighteen-key function keypad, the user could request specific characters, words, sentences, lines, or the entire text to be read. A text-editing mode was facilitated by the system's repeating the text strings entered via a regular keyboard. The highlight of Screen Reader/2 was its access to a significant amount of GUI features. It spoke out the visual effects of text, such as text color, font, and size. It adopted simple non-speech audio to indicate attributes of menu items. It functioned with user controls such as check boxes, buttons, and dialogue windows. Although usability tests were not reported, Screen Reader/2 was the first functionally comprehensive digital information access for visually impaired computer users.

Visually impaired users can now purchase any of the following screen readers: JAWS (Henter, 2003), Window Eyes (GW Micro, Inc, 2003), and Hal (Dolphin Group, 2004), but the cost is much higher than what an individual typically pays for software today. These products are designed to work with Microsoft Windows, Mac OS, and Linux. There is no publication reporting on the usability of these commercial products.

These products inherited and enhanced the functions of Screen Reader/2 by providing more comprehensive, though more complex, keyboard controls and audio outputs.

Although the invention of screen readers is revolutionary, some major issues with these products exist. The navigation and operation capabilities are restricted by the user's ability to memorize complicated keystrokes, the complete set typically consisting of over one hundred variations. Even if these keystrokes are all memorized, a considerable amount of functions and contents are still reached via up and down or left and right arrow keys, which means the user has to go through unwanted information to reach the wanted. The speech output is flat – the hierarchies, indentions, columns, spaces between topics, and other structural information indicating content structure are not presented. Visual effects used to direct users' attention are not conveyed to visually impaired users. The serial presentation of functions and content does not facilitate the user's formation of a structured information space – a pivotal characteristic for information comprehension, thus, losing the richness of the information displayed.

2.2.2 Non-visual Navigation in Hyperspace

Beginning in the mid 1990's, researchers extended their efforts to hyperspace. They created non-visual web page browsers that enabled blind users to join the World Wide Web community. Here, two representative systems are discussed.

2.2.2.1 Lessons Learned from the DAHNI System. Significant work was done by Helen Petrie and her colleagues (Petrie et al. 1996; Petrie et al. 1997; and Morley et al. 1998). They first compared visual interfaces with auditory interfaces, and concluded that visual interface users received a large amount of information at once, which contained

cues supporting the formation of an overview and the choice of content to focus on, whereas auditory interface users did not have this advantage because audio output was transient and serial. Based on their findings the researchers developed a system called DAHNI (Demonstrator of the ACCESS Hypermedia Non-visual Interface).

DAHNI provided tour information of London. The information set contained 37 nodes (i.e., web pages) netted by hyperlinks. DAHNI presented three types of information overview, including a short description of a web page, a scan of the links on the page, and the location of the page using a number assigned to each page (i.e., node). Users could control reading options, reading pace and the settings of the auditory output. The device could accept input from a standard keyboard, a joystick, or a customized touch-tablet. These input modalities were not integrated into a multimodal dialogue, but used individually each time the program was started. The system could accept input from different devices because the congregation of input commands was laid out on an "H" shaped working space and this working space was mapped to each input device.

The auditory output of DAHNI was designed to reflect the hyper-spatial features. When a hyperlink was read, an earcon (i.e., a non-speech auditory cue, functioning similar to an icon (Brewster et al. 1996) was used to mark the beginning of the link without slowing down the speech output. The link was also read in a higher pitch. The use of these combined auditory cues was deemed useful for users' navigation in a system evaluation. A "select" command was accessible on the keyboard for going to the linked page. Earcons were used to indicate headings on the web page and error messages. Tactile output was used in combination with speech description to display pictures.

A comprehensive system evaluation was conducted among nine visually impaired users. Objective task performance and errors (e.g., methods used, time taken, use of the input device, problems encountered) were analyzed using videos, experimenter's notes, and computer logs. Subjective ratings of the users were collected using 5-point rating questions inquiring about all aspects of the system, including the usability of the input devices and commands, the memorizability of the earcons, the usability of system help and the tutorial, the presentation of information, and the feeling of orientation in the hyperspace. Separate experiments were conducted to assess users' learning and memory (e.g., recognition of earcon-indicated headings, recognition of sounds used, and free recall of information presented). Besides the overall positive results of the system evaluation, the keyboard and the touch-tablet were found to be the favored interaction modality over the joystick.

In general, the participatory design approach, the functionality of the system, and the evaluation methods of this series of studies serve as valuable references for later studies conducted on choice of input mechanisms for visually impaired users.

2.2.2.2 User Interface of a Home Page Reader. Home Page Reader is another non-visual Internet browser adopting keyboard input and auditory output (Asakawa and Itoh, 1998; IBM 1998-2004). Besides functions similar to the DAHNI system, Home Page Reader has the following enhanced features: searching for strings, canceling a connection to a new page, moving between pages in a history list, inputting a URL, searching the Web, managing bookmarks, and playing plug-in multimedia files located on a Web page. An email system allowing sending, receiving and composing emails is also built into the browser.

User controls are based on HTML tags on a web page. By capturing the tags of web page elements (e.g., <h1> and <h2> for headers, <a href ...> for hyperlinks, <img ...> for graphs, <table> for tables, etc.) the system allows users to browse either sequentially or by elements and to move forward or backward within the same type of elements. Auditory cues are used to indicate the type of element (e.g., a higher pitch is used when reading a hyperlink). Users perform over 100 functions using 21 keys on either a customized keypad or a standard keyboard. While the functionality is significantly enhanced, the complex keystroke operations make it a challenge to become proficient with the system's full range of functions. In addition, much of this functionality must be taught and cannot be learned by interacting with the system.

Details of the usability test were not reported, but the researchers reported the results of a study used to measure the time needed to adequately train users to beginner, intermediate, and advanced levels. System novices needed additional training about the basic concepts of the Internet such as what homepages and hyperlinks are. Help information needed special tailoring for these novices. Intermediate users needed 30 minutes of in-person training to learn the basic functions. Only advanced users (those with considerable computer expertise with other systems) were asked to learn the advanced functions. Using the online manual and the system's help function, these users were able to learn the system on their own. After one day they were able to use the basic functions and after three days they could use the advanced functions.

To improve their design, the researchers ran an additional study to uncover the maximum listening speeds of the visually impaired (Asakawa et al. 2003). They found the highest listening speed for advanced users was about 500 words per minute and that

for novice users, about 300 words per minute. Both are much faster than the default reading speed of most screen readers and Internet browsers.

The researchers also found that web page designers tend to fragment contents using visual effects (such as background colors and borders). The visually fragmented groupings are not accessible using tag order reading. To solve this problem, the researchers designed techniques to transcode visual effects into structured annotations accessible by information readers (Asakawa and Takagi, 2000, Asakawa et al. 2002, and Takagi et al. 2002).

Non-visual Internet browsers are a significant step toward bridging the gap between visually impaired users and electronic information resources. On the other hand, issues remaining in the interaction methods of these systems are similar to those of screen readers. The amount of information and functions presented sequentially slows down the information access process and provides limited assistance in establishing a mental picture of the information organization. Gestalt information (e.g., proximity, similarity, continuity, closure, etc. of information objects) that sighted users obtain at a glance from a table of contents is not presented effectively or at all. Affordance of interface objects is rarely presented. In short, blind users lack the opportunity that sighted people are given to effectively process information presented to them.

2.2.2.3. Web Page Summaries for Visually Impaired Web Surfers. In a more recent attempt, researchers implemented "gist" summaries for web pages to assist visually impaired users' web browsing (Harper and Patel, 2005). The problem the researchers attempted to solve with available web browsing products, was that of requiring visually impaired users to listen to an entire web-page before understanding its

usefulness for their current task. This sequential access only mechanism handicapped users. "Gist" summaries have also proved useful to sighted users' web browsing. By receiving the same type of summaries, visually impaired users can decide more quickly on the utility of a web page.

A "gist" summary creation tool, "Summate" (Chen, 1997; Chi, Pirolli, Chen, and Pitkow, 2001), was installed to work with the FireFox web browser for visually impaired users. Summate is a client-side system that automatically and dynamically annotates web pages with a small summary at its head. The gist summaries were created "on-the-fly" using the following rules: based on its algorithm, a maximum of four sentences were returned – the first sentence of the web page, the first sentence of the last paragraph, and the sentences at the upper (75%) and the lower (25%) quartiles of the web page. These sentences together were evaluated for their "goodness" as the summary for the page. The measure of goodness was annotated using "high", "medium", or "low". The summary and the denotation were displayed together to the user as a JavaScript generated FireFox alert. A casual test with sighted users on random real web sites confirmed users' preference for gist summaries generated using this algorithm.

2.2.3 Designs Reading Specific Types of Information

Screen readers often do not read special information structures correctly. These information types include tables, mathematical notations, and graphs. Systems reading these information types have been created.

Oogane and Asakawa (1998) first worked on the accessibility of tables in HTML files. They created an index to every individual table cell. Users can use either the table

index or the arrow keys to read table cells. When tables are not arranged in regular rows and columns (e.g., when the first column has one row and the second column has three rows), the tables are transformed to comport with a regular row and column structure.

Yesilada and his colleagues (2004) investigated the method to render tables into audio. They enabled three levels of table navigation in their auditory table browser, EVITA: the low-level navigation of moving left, right, up and down and et cetera, the high-level navigation of moving to designated rows and columns, and comparison between rows and columns. The user controls are mapped onto numeric keys on the keyboard. Users are able to control information flow, as well as to choose what to read next in a table.

A UMA (Univeral Mathematics Accessibility) system is being developed through a multi-institution collaboration (Karshmer et al. 2004). The system converts mathematical documents transcribed in formats used by sighted individuals to those used by unsighted individuals and vice versa. Through the use of the FreeTTS speech synthesis engine and the Java Speech API, the system can render math notations to audio output.

Edwards and colleagues (2006) complemented the above creations by implementing Lambda, a multimodal math editor that presents math notations through Braille and synthetic speech in a linear fashion. Longitudinal observatory evaluations revealed that, instead of using the multimodal output, all testers liked to use the Braille output only but kept the speech output off. To break the sequential access to elements in a math notation, the users created shortcuts to access certain parts of a math notation quickly.

In addition to math notations, access to graphical representations has been researched intensively too. Graphs are presented by either summaries in speech and non-speech audio (e.g., Asakawa and Itoh, 1998; and Zhao et al. 2004), in combined audio and haptic output (e.g., Petrie et al. 1997; Colwell et al. 1998; and McGee et al. 2000), or in tactile output (e.g., Petrie, Morley, and Weber, 1995; Wall and Brewster, 2006). The idea of haptic output is to use the movements of a device or the popping-up of a collection of small pins to generate dynamic outputs interpretable by the user. The idea of tactile output is to allow the user to detect the shape of the surface of the output device (i.e., Braille output). Haptic and tactile outputs have great potential to present graphical information.

The issues of sequential output caused by using left, right, up and down keys take place again in some systems. Alternative input methods are needed to allow direct access to system functions and information items. This means not only information searching based on the relative position of information items, but also direct access based on absolute position should be provided as complementary input methods.

2.2.4 Enrichment of Information Representation Using Audio

Extensive work has been done to enrich information presentation using audios. For the sake of brevity, only a few recent studies are mentioned below.

Brewster and his colleagues used non-speech cues, or earcons, to present the positions of information nodes in a hierarchical information space (Brewster et al. 1996; Brewster, 1998). The hierarchy was formed by general categories, sub-categories, and leaf information items. The timbres, pitches and rhythms of musical earcons were

manipulated to represent different information categories and levels. It was reported that the users were able to comprehend the structural information presented in earcons.

Smith and her colleagues extended the work by using earcons to present hierarchical relationship of objects in an integrated development environment (IDE) for blind computer programmers (Smith et al. 2004). A repetitive simple earcon was used to represent the depth of an information node on the hierarchy – the more repetitions, the deeper the information node.

Cohen and colleagues (Cohen et al. 2005 and 2006) expanded the work by using auditory cues to support navigation and understanding of relational graphs. In relational graphs, nodes are connected by connection lines. Variations of pitch, the use of musical scales, and insertion of audio effects were experimented to represent nodes and connection lines. The final implementation is using a continuous musical tone to indicate that a connection line is being navigated, an increased volume to indicate the departure from the connection line during navigation, and a tone with vibrato effect to indicate the proximity to a node.

Sounds have also been reported as effective representations for the locations of states on an American map (Zhao et al. 2004 and 2005), for the preview of a web page following a hyperlink (Parente, 2004), for two-dimensional tabular numerical information (Ramloll et al. 2001), and for three-dimensional interactive environments for visually impaired children to learn orientation, geography and culture (Sanchez and Saenz, 2005; Sanchez and Baloian, 2005).

Providing richer and necessary information only through audio output has been the focus of researchers in this field. However, well designed audio output is only one

aspect of an effective and intuitive interaction. There is a need to make both aspects, i.e., audio output and user input actions, coupled tightly to form an interaction flow that helps users establish a mental model of the information space through interface dynamics.

2.2.5 Summary of Designs for the Visually Impaired

Through the review of technical products designed for visually impaired users, two general design issues are identified. The first is no ease in learning. Current designs, most of which use either keyboard input or Braille input, require tremendous training and memorization to use. The high learning requirement becomes a barrier for a significant amount of visually impaired users. For people who lost vision in their older age, learning new technical skills and Braille is a much higher challenge. Unfortunately, statistical reports indicate that health problems accompanying aging are the highest reason causing vision lost (American Foundation for the Blind, 2001). The second design issue is sequential access to information. Many system designs require information to be accessed using arrow keys, and as such, users need to go through a large amount of unwanted information to reach the desired information.

To provide solutions to these issues, a new input mechanism needs to be invented. This input mechanism should reduce the learning and memory load for using it and provide direct access rather than sequential access to information contents and system commands. For these purposes, we chose to integrate gesture input and speech input into a non-visual multimodal interaction.

Unlike keystroke input, gesture and speech input are natural and familiar to most people. Speech can provide direct access to information content and system commands

without sequential skipping. Gesture input, coupled with carefully designed audio output, can reduce the learning and memorization load.

In the next section, work related to designing such a multimodal input dialogue is discussed. Since the multimodal input design for the visually impaired is still rare, the main body of the discussion will be based on designs of gesture input, designs of speech dialogues, and designs of multimodal input on graphical user interfaces.

In the following Table 2.1, representative systems for the visually impaired discussed in each category are summarized, along with their strengths and deficiencies.

Table 2.1 Summary of Non-Visual Information Access on Desktop Computers

Technical Products	Screen Readers		Internet Browsers		Browsers Accessing Special Info Types		Products with Specialized Audio Output		
System Example	JAWS	Window-Eyes	DAHNI	Home Page Reader	EVITA	UMA (Universal Math Access)	JavaSpeak (A plug-in based on Eclipse Platform)	Sonification of geo-referenced information	Audio enriched links
Application Domain	GUI access and Internet support	GUI access and Internet support	Internet access	Mainly Internet and email access	Table navigation	Math notation access	Non-visual Java programming	Geo-info access on a U.S. map	Web page previews
Developer	Freedom Scientific	GW Micro, Inc.	University of Hertfordshire, UK	IBM	University of Manchester	Multi Univ. project	Multi Univ. project	Univ. of Maryland, College Park	University of North Carolina
Reference	Developer's web site	Developer's web site	Morley et al. 1998	Asakawa and Itoh, 1998	Yesilada et al. 2004	Karshmer et al. 2004	Smith et al. 2004	Zhao et al. 2004	Parente 2004
User Input / System output	Keyboard / Synthetic speech or Braille	Keyboard / Synthetic speech or Braille	Keyboard, joystick or touchtablet / Audio, audio plus tactile for pictures	Keyboard or customized keypad / Audio	Keyboard / Audio	Keyboard / Audio	Keyboard / Audio	Keyboard / Audio	Keyboard / Synthetic speech
Strength	Rich functions for Windows GUI access	Rich functions for Windows GUI access	① Finely designed audio output features for navigating the hyper-space ② Logical layout of commands on a user input "workspace" usable by multiple input devices	Rich functions for Internet access	① Non-linear table browsing enabled ② Independent table browsing as well as table linearizing for read by screen readers	Conversions between various math notations used by sighted users and visually impaired users	Finely designed audio output for representing hierarchical structures	Spatial sounds representing location of a state on a mosaic U.S. map	Web page preview before following a hyper link
Challenges to Users / Deficiencies	① Considerable amount of learning & memorization to use via keystrokes ② Braille display is expensive ③ Tables are read in a linear manner	① Considerable amount of learning & memorization to use via keystrokes ② Braille display is expensive	Skip of unwanted information only in a sequential manner by using "next" and "previous" like commands	Considerable amount of memorization to use over 100 system features via 21 keys	Usability study only done with one blind user. Report not included how the blind user understands tables and whether s/he needs to know spatial locations of cells	Usability study not yet reported	Sequential skip of unwanted information by using arrow keys	Sequential browsing of all states from top to bottom column by column	Compatibility only with Internet Explorer used with conjunction of JAWS

2.3 Designs of Speech and Gesture Interaction

Linguists have recognized the close synchrony between gesture and speech in human communication. They recognized the synchrony from the fact that when speech was disrupted gesture was disrupted too (McNeill, 1992), that stutterers modified their gestures to match with their speech (McNeill, 2000), and that deliberate mismatch between gesture and speech could influence a subject's recall of a narration (McNeill, 1992). The synchrony indicates that gesture and speech are inseparable units expressing different aspects (i.e., the imagistic aspect and the linguistic aspect) of the same conceptual content in the communication (McNeill Lab). This synchrony suggests that combining hand gestures and speech input in the human-computer interaction not only adapts to the natural way of human communications but also provides potential powers in expression.

Currently, little work is available on multimodal hand gesture and speech input dialogues on non-visual interfaces, but such dialogues have been studied on visual interfaces. These studies, together with trails in designing hand gesture input alone and speech input alone, provide indications on designing non-visual multimodal dialogues and nurture the establishment of a theory base. In this section, the following related aspects are discussed: (1) designs of gesture input, (2) designs of speech dialogues, and (3) findings and interpretations from existing studies on multimodal hand gesture and speech interface dialogues.

2.3.1 Gesture Input Design

Well designed gesture input can be intuitive to blind users. It has better capability to imitate direct manipulation than keypad/keyboard input, the main stream of input devices for the current devices for the blind. Because direct manipulation is a key feature of intuitive interaction with WIMP systems (Windows, Icons, Menus, Pointer), gesture input has advantages in providing powerful access to WIMP applications.

2.3.1.1 Classification of Gesture Input. Hand gesture input is one of the most studied input methods. When designing gesture input grammars, designers refer to natural gestures used in human communication. Human communication gestures can be categorized as manipulative gestures, semaphoric gestures, and pointing gestures (Quek et al. 2002). Manipulative gestures intend to control some entity by applying a tight relationship between the actual movements of the gesturing hand/arm with the entity being manipulated. Grasp, release, drop are typical manipulative gestures. On graphical user interfaces, manipulative gestures are especially used in direct manipulation, and may be aided by visual, tactile, or force-feedback from the object (virtual or visual). Semaphoric gestures employ a stylized dictionary of static or dynamic hand or arm gestures. An example is moving toward right to indicate moving forward. Semanphoric gestures are communicative in that they serve as universal symbols for the human-machine communication. Pointing gestures are mainly used for deictic purposes in combination with definite articles or demonstrative pronouns.

Manipulative gestures have been used in virtual reality games to assist stroke rehabilitation (Merians et al. 2002; Boian et al. 2002). The patients' gestures are collected using sensory gloves. The patients' tasks are to manipulate virtual objects including a

window, a butterfly, a piano, and several pistons. The range of the hand motion is exercised by wiping the glass of a window to see the landscape outside. The speed of the hand motion is exercised by scaring away the butterfly. The fractionation of the fingers is exercised by playing the piano. The strength of the finger is exercised by moving the piston connected to each finger. A three-week patient trial was conducted among four post-stroke hemiplegic subjects aged 58 to 72. The results showed various degrees of improvement in hand impairment following this therapy. There was a good retention of gains and a positive subjective evaluation by the patients and participated therapist.

Semaphoric gestures are broadly used for text input (e.g., Graffiti) and for conveying geometric attributes. Graffiti is a single stroke alphabet that resembles the Roman alphabet and is based on Unistroke (Goldberg and Richardson, 1993; Költringer and Grechenig, 2004). When Graffiti input is used, the user's hand gesture input is recognized and transformed to letters, numbers, backspace or special characters. Hand gestures are usually transmitted to the computing device interface through a stylus because of the limited physical space for conducting Graffiti input. In various studies Graffiti has proved to be an efficient text entry method (Fleetwood et al. 2002; MacKenzie and Zhang, 1997).

Sowa and Wachsmuth described a system that used co-verbal iconic gestures for describing objects in a virtual environment (Sowa and Wachsmuth, 1999 and 2000). In the study the subjects described a set of five virtual parts (e.g., screws and bars) that are presented visually to them in wall-size display. Their descriptions were in combined speech and gesture, which were captured by a microphone and a pair of CyberGloves (Immersion Corporation). The researchers found that the subjects presented geometric

attributes by abstracting parts from the complete shape using semaphoric gestures, e.g., using combinations of movement trajectories, hand distances, hand apertures, palm orientations, hand-shapes, and index finger direction.

Pointing gestures lie in between manipulative gestures and semaphoric gestures. Pointing gestures are mainly used as deixis in multimodal input in combination with verb instructive commands. A well-known instantiation is Bolt's "Put-That-There" system prototype with "that" and "there" instructed using pointing gestures (1980). The advantage of such combinations is their adaptation to the natural human communication patterns.

Several design strategies can increase the usability of these gesture inputs. The first is using real spatial references (such as Badler's plastic spaceship (Badler et al. 1986)) as opposed to an imaginary objects for gesture manipulation, because "[locating] a desired point or area [is] much easier when a real object is sitting on the Polhemus's digitizing surface." (Ostby, 1986) The second is applying physical constraints (Norman, 1990). Software constraints, although useful, often require the understanding of the constraints and their feedback, which impose a small cognitive load. Physical constraints can lend support and remove this cognitive load: users can try configurations of objects by moving their fingers until they hit something (Hinckley et al. 1994a). The third is implementing gesture manipulations onto a small working space to comport with the typical small physical working volume exhibited by users. This user behavior was seen in observations of subjects performing writing tasks (Guiard, 1987) as well as observations of users' performance on a two-handed gesture interface (Hinckley et al. 1994b). These

design strategies are valuable and have been reflected in the design of the hand input of our experiment system, AudioBrowser.

2.3.1.2 Gesture Input Designs Influential to Non-Visual Interfaces. An array of innovative interaction mechanisms has lighted made great strides in a hope to helping to solve the design issues in designs for the visually impaired users. These interaction mechanisms take advantage of gesture and tactile inputs, whose potential to affect present designs for the blind is prominent, but has not yet been fully recognized.

Roth and his colleagues created an interface that receives user input from a touch screen and generates 3D auditory output for the user (Roth et al. 1998; Roth et al. 2000). When a user points to a block of content on a web page displayed on the touch screen, the content is outputted via speech with added spatial characteristics that help characterize the location of the information. For example, if the information is on the top-left on of the screen, the speech output appears to come from the top-left portion of the user's hearing space. This location information can provide navigational assistance for later revisit subsequent visits of the same content.

Brewster and his colleagues developed an eyes-free gesture-audio interface for wearable devices (Brewster et al. 2003). The interface receives input from the user's head movements and hand gestures and produces 3D auditory output. The head gestures are received via a head gesture detector mounted on the user's headphone. The hand gestures are received via the touch screen of a PDA. These gesture inputs proved to be effective even when users were walking. However, the head gesture detection device is very costly. Moreover, the head gesture may be awkward when used in public. These types of

gross motor input mechanisms also do not allow for a large array of options to be input to the computer by the user.

Friedlander and his colleagues created Bullseye menus on touchpads (Friedlander et al. 1998). The Bullseye menus consist of a set of concentric circles divided into quadrants. A menu item is located in each quadrant. Non-speech audio cues are used to indicate the boundaries between menu items and the direction of the stroke (up down, left, right) further extends the number of menu items that can be accommodated with this method. In the evaluation of Bullseye menus as a potential input mechanism for visually impaired users, it was found that users could efficiently and effectively select a large number of menu items using non-speech audio feedback. A second study looked at tactile feedback, which gave a slight "bump" as each concentric ring was passed. This, too, worked effectively, but was not as efficient as the sound feedback.

A variation of the Bullseye, called the earPod, was presented recently by Zhao et al. (2007). The earPod is an eyes-free menu selection interface using touch input and reactive audio feedback. Up to 12 menu items can be mapped on its circular track. Browsing menu items is executed by sliding the thumb on the circular touchpad. Selection is executed by lifting the thumb from the desired menu item. When a menu item is touched, speech output is synthesized to read the item. When a boundary between two menu items is reached, a click sound is made. When finger motion is fast, audio feedback is truncated to give partial playback. A post-evaluation showed that there wasn't significant difference in the selection accuracy and overall selection speed between the earPod interface and its counterpart interface with only visual feedback. Half

of the sighted participants preferred the earPod interface while the other half preferred the visual interface.

A series of work has been conducted by Tremaine and her co-workers (William and Tremaine, 2001; Chen et al. 2003, 2004, 2005 and 2006). They created a mechanism using simple and intuitive pointing gesture and button clicks on the touchpad to control information browsing. Soundnews (Williams and Tremaine, 2001), an early version of this interface, was used to browse hierarchically organized news articles on desktop computers. This mechanism was then implemented on a PDA interface, the AudioBrowser system, to access personal information (Chen et al. 2004, 2005 and 2006). The idea is dividing the sensing area of the touchpad to small segments that are mapped with information items and operation commands. When a segment is touched, the system speaks aloud the information item or operation that is touched. The user then clicks the buttons on the touchpad to execute the command, or to zoom into the detailed information in which case the segments on the sensing area are changed accordingly to map with the information items on the new information hierarchy. This mechanism relieves the user from having to memorize operation vocabularies and supports exploratory learning. This mechanism also provides an advantage similar to that of a visual interface wherein the user is able to browse the operation options for possible future use. The prerequisite to use this mechanism is that the information to browse is organized hierarchically. The usability studies on both versions proved that the users made efficient use of this mechanism.

Text input has been impossible on non-visual interactions by users who do not know Braille input or regular keyboard input. But Wobbrock's EdgeWrite© makes non-

visual text entry possible for those users (Wobbrock et al. 2003; Wobbrock et al. 2004). EdgeWrite was originally designed for people with motor disabilities. It is similar to Graffiti text entry mechanism but relies on physical edges and corners of the input device (e.g., the sensing area of a touchpad). The user moves his or her stylus or finger along the physical edges and into the corners of a square bounding the input area. Recognition of the user input does not depend on the path of the motion, but on the order that corners are contacted. EdgeWrite was first implemented on a Synaptics touchpad. Recently it has been customized for a four-key keypad (Wobbrock et al. 2006), trackballs (Wobbrock and Myers, 2006), and joysticks on cell phones (Wobbrock et al. 2007) and wheelchairs (Wobbrock et al. 2005). Because of its ease of use and its suitable input device, EdgeWrite on the touchpad is very promising for non-visual text entry.

Another invention that has potential impact on mobile text entry is the non-keyboard QWERTY typing developed by Goldstein and colleagues (Goldstein et al. 1999). This typing mechanism uses pressure sensors strapped on fingertips to detect pressing motions of fingers. A language model based on lexical and syntactic knowledge is used to transform finger stroke sequences into words and sentences. A keyboard is no longer a necessity necessary for using the QWERTY input. This text entry mechanism, although still requiring users' familiarity to keyboard input, allows fast input while maintaining the mobility of the wired device.

Based on the review above, advantages of gestural input can be summarized as the follows. First, the prevalent use of gestures in human communications makes intuitive gesture input on computers possible. Gesture input commands can be designed so that traditional gesture meanings in human communications are carried on (e.g., Brewster et

al. 2003). Second, this type of input can provide advantages similar to those of direct manipulation on GUIs, and can reduce the demand of commands recalled (e.g., Roth et al. 2000, Friedlander et al. 1998). And third, when used in conjunction with sensor technologies, portability of the connected device can be achieved (e.g., Wobbrock et al. 2003 and 2004, Goldstein et al. 1999). These advantages target the same design issues in systems for visually impaired users recognized in previous sections.

2.3.2 Speech Input Dialogue Design

Conversational speech offers an attractive alternative to input methods on physical spaces. It is familiar, requires minimal physical effort for the user, and leaves the hands and eyes free. Since speech is not constrained by physical dimensions, the number of speech commands is virtually unlimited.

Spoken dialogues can be designed simple or complex depending on the degree of freedom the user is given and the naturalness of the conversation. A definition given by Fraser (1997) described spoken dialogue systems as computer systems wherein humans interact on a turn-by-turn basis and in which natural speaking plays an important part in the communication. Spoken dialogue systems allow users to interact with complex computer applications in a natural way using speech.

2.3.2.1 Classification of Speech Dialogues Based on Dialogue Control Strategies. Speech input dialogues can be classified based on three dialogue control strategies (McTear, 2002):

Finite-state dialogues: A finite-state dialogue consists of a sequence of predetermined steps. In most cases the system has complete control over the

conversation, produces prompts at each step, recognizes user entries, and produces system actions based on user entries. The user's speech entry is usually short and must adhere to predetermined grammar that is carefully prompted by the system. A major advantage of this type of dialogue is its simplicity and relatively high accuracy rate of speech recognition. The vocabulary and grammar specified in advance ensures less recognition errors. However, the disadvantage is its lack of flexibility and naturalness. The user is constrained either to input one value at a time or to input multiple values according to a strictly defined form and order.

Template-based (or frame-based) dialogues: Rather than building a dialogue according to a predetermined sequence of steps, a template-based (or frame-based) dialogue analogizes a form-filling task in which a predetermined set of information is to be gathered. The user provides required information in a flexible order. The system fills the provided information into the predefined template, and prompts for any missed information items. Similar to state-based dialogues, template-based dialogues are suitable for well-structured tasks, and in most cases the system takes the initiative and elicits data from the user to complete a task. The difference from a finite-state dialogue is that the flow of a template-based dialogue can be event-driven and not predetermined. Template-based dialogues allow more flexible and more natural user entries.

Agent-assisted dialogues: The third type of dialogues is assisted by one or more intelligent agents. An agent cooperates with other agents if it is unable to handle the task alone. Particular subtasks may be assigned to particular agents. Communication among agents is monitored and used to generate a state-of-play report that is further used to supply information required by the other agents and arrange the interaction activities.

With this type of dialogues, the system is able to intelligently and flexibly solve most complex tasks for the users.

The spoken dialogues can be led by the system, the user, or both. In system-led dialogues the system asks the user a series of questions to collect information required to fulfill a task. The system determines what questions to ask and in which order. System-led dialogues are usually modeled into a state transition network or a transition diagram (Green, 1986), in which the nodes represent the system's questions and the transitions between the nodes determine all the possible paths through the network. In user-led dialogues, the user initiates the conversation and system actions. In dialogues led by both the user and the system, a rule-based expert system may be needed to manage complex conversations based on decisions rules and contexts.

A study conducted by Potjer and his colleagues (Potjer et al. 1996) compared a system-led version and a mixed-initiative version of a simple call assistance application. The system-led version used isolated word recognition. The mixed-initiative version used continuous speech recognition and more complex natural language processing. When using the system-led version, the user provided the required information in two steps; when using the mixed-initiative version, the user provided the required information in one utterance. However, the results from a performance comparison indicate that the system-led version was not slower than the mixed-initiative version, because the mixed-initiative version encountered more recognition errors due to multiple meanings embedded in longer sentences.

In terms of user satisfaction towards different types of dialogue controls, Potjer's research (Potjer et al. 1996) showed no difference between the system-led interface and

its mixed-led counterpart. Billi and his colleagues (Billi et al. 1996) verified that dialogues that required the user to input single words or simple phrases achieved a better user acceptance than dialogues requiring more complicated user speech input.

These results were extended by a study on a multimodal gesture and speech input interface by Robbe-Reiter and co-workers (Robbe-Reiter et al. 2000). Reiter's study indicated that compared to compound speech input, simple and short utterances reduced neither the efficiency of users' interactions nor their satisfaction. In addition, the fixed grammar on short utterances had a limited influence on users' use of modalities.

To guide developers' choice of speech dialogue control strategies, Bernsen and Dybkjærs (1994) created a task-oriented dialogue theory stating that small and simple tasks should use single-word dialogues, larger and well structured tasks in a limited application domain should take use of system-directed dialogues, and that larger and unstructured tasks should use mixed-initiative dialogues.

Since the tasks are relatively simple in the AudioBrowser system, we implemented user-led finite-state dialogues. Speech commands are simple words and phrases. Compared to other alternatives, this type of voice interaction is the simplest and hence, easier to learn and use, and achieves a better speech recognition rate.

2.3.2.2 Lessons Learned from Speech Dialogue Designs. For many reasons, speech recognizers are still error-prone. Also, opposite to graphical interface, a speech-only interface makes system functionality and operation boundaries invisible (Yankelovich and Lai, 1998; Yankelovich, 1996). As such, speech-based interfaces should emphasize feedback and verification to guide users through a successful interaction. A list of prompt

and confirmation techniques has been indicated by researchers (Brennan and Hulteen, 1995; Yankelovich et al. 1995; Yankelovich, 1996).

Significant lessons can be learned from Yankelovich's study series (Yankelovich et al. 1995; Yankelovich, 1998). The system studied was an experimental conversational speech system that provided mail, calendar, weather, and stock quotes applications. The user studies produced several design indications. (1) Without visual aids, the speech interaction is more like interpersonal conversational style and away from graphical techniques. GUI conventions would not transfer successfully to a speech-only environment. Vocabulary on the graphical interface was not the same as that in speech interactions. For example, relative dates such as "one week from tomorrow" and "the day after Labor Day" were not necessary on a graphical calendar but were frequently needed by users of a non-visual speech interface. (2) Information flow controls such as pop-up dialog boxes that worked efficiently in GUI were highly non-compliant by users of the non-visual speech interface, suggesting that dialog boxes should be eliminated. Auditory feedback should be carefully designed to announce and confirm the dialogue modes change. (3) Speech was easy for humans to produce, but much harder for us to consume. The slowness of the speech output was a main contributor. As such, the speech feedback should be informative, relevant, brief, orderly, and no more than is required. Systems should permit users to move the conversation forward more quickly by accepting compound answers (A discourse management module in the program should keep track of what has been said and what has not (Martin et al. 1996)). Users should be able to interrupt the output with their voice.

The results of Yankelovich's studies reached a consensus with Mynatt's conclusion (Mynatt, 1994; Mynatt and Edwards, 1992 and 1995; Edwards and Mynatt, 1994), that when a GUI is transferred to a non-visual interface, the non-visual interface should not model every aspect of its visual counterpart. The non-visual interface should present its content structures differently to reflect users' specific need for navigation in a non-visual condition through auditory presentation media. Furthermore, even if the interface keeps its GUI features, when speech input is added, the GUI feedback should be redesigned to allow users to be aware and to use the full potential of the speech dialogues (Ibrahim and Johansson, 2002).

The lessons and principles in design of voice input dialogues should benefit the design of non-visual multimodal interface dialogues that involve speech input techniques.

2.3.3. Multimodal Interaction Design

Multimodal speech and gesture input dialogues are not yet available in technical products or published research. The majority of existing studies on multimodal interactions are on GUIs. Studies providing insights on behavioral patterns of user inputs have been mainly conducted within two application domains: interactive map related tasks and multimodal disambiguation and error correction.

The map-based user studies were conducted on interactive simulated maps using Wizard of Oz user study methods. Users of these maps used stylus or mouse to point or circle an area on the map, gave natural speech instructions (such as "I don't want houses in the flood zone", or gave commands in written words or symbols (such as a cross).

Based on the user's input captured, a hidden experimenter sent predefined response interfaces to the user's terminal.

Multimodal disambiguation and error correction were observed in a spectrum of multimodal systems, and especially on speech recognition systems. These systems were large vocabulary speech recognition systems such as ViaVoice ® (IBM) and Dragon Naturally Speaking (ScanSoft). These systems allow users to manually correct speech recognition errors using speech commands, mouse selection, and keyboard typing.

In the following subsections, the advantages of combined gesture input and speech input are first discussed. They are followed by the review of findings and interpretations of user behaviors in the two multimodal application domains, i.e., interactive map related tasks and multimodal error correction. Although there is an array of technical multimodal products created and presented from the technical standpoint in the literature, those products are not discussed in this review, since this review is to provide references for a dissertation research focusing on the user behavior.

2.3.3.1 Expanded Powers of Combined Gesture and Speech Inputs. Combined gesture and speech have been documented to have various advantages in terms of expressiveness, complementariness, robustness, and performance efficiency on computer interfaces.

Expressiveness: Combined gesture and speech inputs allow more powerful expressions. Gesture and speech are both semantically rich input modes, but they have been observed to have different expressive powers in describing subjects in different domains. Gestures are powerful in describing geometrical attributes because they provide several methods of expression that speech does not provide, such as combinations of

movement trajectories, hand distances, hand apertures, palm orientations, hand shapes and index finger directions (Sowa and Wachsmuth, 1999 and 2000). On the other hand, speech provides a complicated vocabulary to address nominal information or instructional commands (Cohen 1992) and is not tied to spatial constraints. Hence, speech can be used to interact with the system regardless of degree of visual exposure (Billinghurst 1998). When speech and gesture are used together, the strengths of each input mode compensate for the weaknesses of the other. Due to the increased expressiveness, users preferred combined speech and gesture input to either modality alone. Hauptman and MacAvinney (1993) found that when a combined input was possible subjects used combined speech and gesture at 71% of the time as opposed to speech input only or gesture input only. Oviatt (1996) found that when users interacted with a graphical map, the more spatial the task was the more users preferred integrated speech and gesture input to either speech or gesture input alone.

Complementariness: Speech and gesture are used in a complementary manner in both human natural communication and human-computer interactions. Linguists have documented that spontaneous speech and gesture do not involve duplicate information (Cassell et al. 1994; McNeill 1992). Oviatt and her colleagues (Oviatt, 1996 and 1997; Oviatt, et al. 1997) found that when users used an electronic graphical map, speech input and pen input consistently contributed different and complementary semantic information – that subject, verb, and object of a sentence were usually spoken, and locative information written. Even when correcting system errors speech and pen inputs rarely express redundant information. Complementary use of speech and gesture is the natural and dominant theme during users' multimodal interaction.

Robustness: From a usability standpoint, multiple input modalities offer an error-handling advantage so that when one input mode fails the user can switch to another. For example, integrated natural language with direct manipulation overcame limitations of the two input techniques when used alone (Cohen et al. 1989). Users select the input mode that they think is less error prone for a particular task, which leads to error avoidance (Oviatt and Cohen 2000, Suhm et al. 2001). To offer the full error-handling advantage the system design needs to meet two points. One is that different input modes provide parallel or duplicate functionality, in order for the user to switch between input methods freely at any point. The other is that different input modes can disambiguate each other so that recognition errors from unimodal recognition can be recovered (Oviatt, 1999a). The results of Oviatt's study (Oviatt, 1999b) showed that one out of eight commands processed by the multimodal system produced the correct response because of mutual disambiguation between the speech and the pen inputs. Oviatt (1996) also found that integrated speech and pen input produced 36 percent fewer task errors.

Performance Efficiency: From a psychological standpoint, users' efficiency of performing multiple tasks can be increased if the tasks require information processing by different sensory modes, for example spatial and verbal (Treisman, 1973). In multimodal interfaces users can perform visuo/spatial tasks at the same time as giving verbal commands with little cognitive interference (Billinghurst, 1998). Martin (1989) found that people using a CAD program with the addition of speech input had an improved performance by 108 percent than people who used a traditional interface. Oviatt (1996) found that integrated speech and pen input resulted in 23 percent fewer spoken words and 10 percent faster completion time compared to speech input only. Suhm et al. (2001)

confirmed that multimodal error correction for speech user interfaces was faster than unimodal correction by respeaking. Grasso found that the task completion time, the number of speech errors, and user acceptance all improved when an interface comprised of both direct manipulation and speech input (Grasso, Ebert, and Finin, 1999).

2.3.3.2 Findings and Interpretations from Integrated Speech and Pen Inputs on Interactive Maps. Oviatt and her co-workers conducted a series of exploratory analysis of users' use of multimodal input versus single-modal input on a simulated interactive map (Oviatt 1996, and Oviatt et al. 1997). The map was called a "Service Transaction System" on which subjects could display, zoom, select, and filter real estate based on given requirements. During the study, eighteen subjects first received a general orientation on how to enter information on the touch tablet when writing, speaking, and combining both modalities. Subjects were free to use cursive handwriting or printing, gestures, symbols, graphics, pointing, or other marks using the stylus. When speaking, subjects were instructed to tap and hold the stylus on the map as they spoke. In all cases, subjects were encouraged to speak and write naturally using the input modalities in any way. For example, a person might circle a lakeside house icon with the stylus and say "I don't want a house in a flood zone." In response, the system would display waterways and flood zones, and would filter out the house icon if it was located in such zones.

The system responded to subjects' input in a Wizard of Oz approach. An assistant tracked and interpreted user input and sent predefined map displays and confirmations back to the subject. The response delays, as reported, averaged less than 1 second between subject input and system feedback.

The research design was a completely crossed factorial with repeated measures. The main factors are (1) input modalities – speech-only, pen-only, and combined speech and pen input, and (2) presentation format – high-resolution map with detailed display of objects and low-resolution map with minimal objects displayed. Eighteen subjects participated in the study. Each subject completed two tasks in each of the six experiment conditions. Subjects' performance was videotaped and transcribed for analysis.

Results found in the studies were rich. The results are shown in Table 2.2.

Table 2.2 Findings from Oviatt's Studies on Integrated Speech and Pen Inputs on an Interactive Map, "Service Transaction System"

Length and Complexity of Speech Utterances	In these map-based interactions, utterances during multimodal input were significantly briefer and less complex syntactically than those during speech-only input. (4.79 versus 6.22 words respectively)
Spoken Disfluencies	(1) Spoken disfluencies during multimodal input were significantly lower than those during speech-only input. (2) A strong relation was found between the length of the utterance and the likelihood of a spoken disfluency.
Spatial Location Descriptions	(1) Utterances in describing spatial locations in speech-input mode were lengthier than those in multimodal input mode. (2) Hence spoken disfluency rate when describing spatial location was higher in speech-input only mode than in multimodal input mode. (3) Also spoken disfluency rate was significantly more elevated in the spatial location domain than in the verbal commands domain.
Task Completion Time	Task completion times were significantly shorter during multimodal map interactions than during either speech- or writing-input only.
Task-critical Content Errors	Average user content errors during multimodal input were significantly lower than either speech-input or writing-input only.
Self-reported and Observed Preferences	If to choose one input method, 94.5% subjects preferred multimodal input, 5.5% preferred writing-input only, and none preferred speech-input only.
Input-task dependence	Spatial location commands on the map were more frequently expressed multimodally, and spatial location commands were the only ones more likely to be expressed multimodally. Speech was used for 100% of subject and verb constituents.
Linguistic Content	(1) 98% of unimodal spoken consructions and 97% of multimodal constructions were in the format of subject-verb-object. (2) Pen input was mainly used for drawing graphs, symbols, signs, and pointing.
Multimodal Integration Patterns	Most (or 86%) multimodal constructions were draw-and-speak. The rest were point-and-speak. Drawing and speaking could be simultaneous, sequential, and compound.

Following this series of studies, Oviatt and her colleagues conducted another study series on pen-and-speech multimodal integration patterns of subjects with a larger age-span. The researchers modeled the behavior of 12 subjects aging 21-58 (Oviatt et al. 2003) and 15 seniors aging 66-86 (Xiao et al. 2003) on a multimodal map system with a random error generator and two experiment "wizards". The system recognition error rates were controlled to allow researchers to observe the change of users' multimodal integration pattern under the effect of errors.

Results showed that all users had a dominant temporal multimodal integration pattern, i.e., to either give multimodal input simultaneously with a temporal overlap between input signals of different modalities, or give the input sequentially with a temporal lag between the signals. The dominant temporal pattern of a user was reinforced with the elevation of system recognition error rate. When the error rate was increased, the multimodal input signal overlaps of simultaneous integrators, and the multimodal input signal lags of sequential integrators, were all correspondingly elevated.

To observe whether children have such multimodal patterns, a study was conducted with 24 children aged 7-10 (Xiao et al. 2002). Because the map-based tasks were not suitable for children at this age, the study was conducted on a simulated system that taught children about marine animals through graphics and animations. The children interacted with the system using natural speech and pen drawing (Oviatt and Adams, 2000). The results showed that similar to adult users, a child was either a simultaneous integrator or a sequential integrator. Their use of simultaneous integration was more often than adults. The lags between their sequential input signals were shorter than adults. For a

child this temporal multimodal integration pattern was dominant during the whole experiment session.

The second series of studies revealed that users of different ages had predominant temporal multimodal integration patterns, which was not likely to change despite changes in system error rates.

2.3.3.3 Findings and Interpretations from Multimodal Error Correction. Various studies have reported that multimodal user controls benefit error handling in error-prone input modes, especially speech input. Those studies uncovered the following user error handling behaviors during their multimodal interactions.

After learning a multimodal interface, users select the input mode that they think is less error prone for particular task, which leads to error avoidance (Oviatt and Cohen 2000, Sears et al. 2003). For example, when inputting a foreign surname, users are more likely to use writing input rather than speech input because users experienced fewer errors in writing names (Oviatt and Olsen, 1994). When users originally preferred one input modality, they also used other modalities because they learned to avoid ineffective input modalities with experience (Suhm, et al. 2001).

Switching input modality was found to be an effective error handling strategy. When recognition errors occur, users switched input modes for error correction (Oviatt, 1999b). It was reported that the likelihood that users switching input mode following a system error was three times higher than in the baseline condition, in which recognition was error-free (Oviatt et al. 1998).

During a study of detailed user error handling behavior on a speech recognition system that provided alternative input modalities for error correction, Oviatt and VanGent (1996) observed two phenomena. One was that users re-dictated the same word despite incorrect recognition. This was referred as "spirals" or "no-switch" strategy. The number of re-dictation transactions was referred to as "spiral depth". The other was that during error correction a new error occurred, which was referred to as "cascade". In a study on three commercial large vocabulary continuous speech recognition systems, Karat and her co-workers found spirals and cascades embedded in each other, and uncovered similar initial no-switch strategy and a higher likelihood of input method switch if errors were not corrected after the initial no-switch tendency (Karat et al. 1999).

Multiple input modalities result in more efficient and better error correction results. In a study on a semi-simulated electronic map system that accepts speech input and pen inputs, Oviatt (1999b) reported that one out of eight commands processed by the multimodal system produced the correct response because of mutual disambiguation between speech and pen inputs. Oviatt (1996) reported that integrated speech and pen input resulted in 36 percent fewer task errors, 23 percent fewer spoken words, and 10 percent faster completion time compared to speech input only. In a study comparing various unimodal and multimodal error correction strategies on speech recognition systems, Suhm and his colleagues (Suhm et al. 2001) confirmed that cross-modal repair speeded up the correction of speech recognition errors on speech user interfaces comparing to unimodal error correction by re-speaking, and that cross-modal repair was more accurate than unimodal repair. The better error correction results with multimodal input than unimodal input is caused by users that are less productive with ASR

(Automatic Speech Recognition) than with keyboard and mouse on correcting dictation errors (Karat et al. 2000), and that speech is less effective in direction-oriented navigation (e.g., move up two lines) needed in speech error correction tasks than in target-oriented navigation (e.g., select target) (Sears et al. 2003).

Multimodal disambiguation is effective not only in interactive map systems and speech recognition systems, but also in other systems. Holzapfel and his colleagues (2004) found that speech input could disambiguate 3D gesture input that was more error-prone than speech. They found that their multimodal fusion approach was very tolerant against falsely detected pointing gestures.

Because of the uncovered error correction behaviors, Oviatt's suggestions on designing multimodal systems are apparently valuable. She suggested that systems with multimodal input modalities should implement two design tactics. One was that different input modes provide parallel or duplicate functionality in order for the user to switch between input methods freely at any point. The other was that different input modes can disambiguate each other so that recognition errors from unimodal recognition can be recovered (Oviatt 1999a).

In general, multimodal input has a great potential to benefit visually impaired users, because "... well-designed multimodal systems should be able to integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other." (Oviatt, 1999b, pp. 576)

2.4 Theories in Cognitive Psychology Applied to Multimodal Interaction

In order to design multimodal interaction, knowledge in how human attention and working memory are used to process multimodal information is necessary. Established theories can be used to explain humans' ability to process information communicated via multiple modalities separately using distinct cognitive resources, and when needed, in a multi-tasking manner by sharing attentional resources.

Although a large portion of work in cognitive psychology is still assumptional and under exploration, some theories have been supported by empirical information and widely accepted.

2.4.1 Human Attention in Relation to Multimodal Interaction

2.4.1.1 Definition of Attention. The earliest but most concrete definition of attention was given by William James, one of the first major psychologists:

> "Everyone knows what attention is. Focalization, concentration, and consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state". (James, 1890, pp. 403-404)

This definition implies that attention is limited – we can only attend to one thing at a time, and that attention is selective – we direct our attention to one thing or another.

2.4.1.2 Models of Attention. In order to understand human attention, models of attention have been built from both cognitive and clinical perspectives.

One of the most commonly accepted attention models from the cognitive perspective was defined by Wickens (1984). According to this model, information received by receptors passes through three stages: encoding, central processing, and responding. In the perceptual encoding state, information is identified and interpreted. Information is further comprehended, integrated and transformed in the central processing stage. Finally, in the responding stage, actions are taken on the basis of the central processing.

One of the most widely used clinical models was defined by Sohlberg and Mateer (1989). This model was derived from research on the recovering of attention processes in brain damaged patients, and is used to evaluate attention in patients with different neurologic pathologies. The model describes five attention hierarchies: (1) focused attention, referring to the ability to respond to specific stimuli; (2) sustained attention, referring the ability to maintain a consistent response to a continuous stimuli; (3) selective attention, referring to the ability to focusing on a specific process while ignoring others, explaining the cocktail party effect (Arons, 1992); (4) alternating attention, referring to the ability to shift attention focus and move between tasks; and (5) divided attention, which is the highest level of attention that refers to the ability to process multiple tasks at the same time.

2.4.1.3 Ramifications of Attention Allocation Theories. Three major ramifications of attention theories have been established: the bottleneck theory, the single resource pool theory, and the multiple resource pool theory.

The bottleneck theory (Welford, 1952; Broadbent, 1958) or the filter theory as referred to earlier, proposes that information is processed in serial order. Along the

processing stages bottleneck exists somewhere which forces the filtering out of information not selected for further processing. In this theory, the capacity for information processing is considered to be fixed, and the processing is done by a single undifferentiated resource – only a limited amount of information can be brought from the sensory register to the working memory.

The single resource pool theory (Kahneman, 1973) on the contrary proposes flexible central resource capacity. In this model the amount of available attention is determined by the individual (individuals' arousal level of attention varies), the tasks (different tasks demand different levels of attention), and the situation (either involuntary, in which something draws attention, or selective, in which an individual consciously decides to pay attention to something). This theory further suggests that, although being a single pool of resource, attention can be allocated to several activities at once, and that parallel processing can occur in all the processing stages. However, when a certain task demands a high level of attention, performance of other concurrent tasks is degraded.

The multiple resource pool theory (Navon & Gopher, 1979; Wickens, 1980, 1984 & 1992) is the most popular one among the three ramifications. It argues that instead of sharing a single pool of resource, there exist multiple pools of resources, each of which has its limited capacity and is related to specific skill. Multiple tasks can be performed at the same time as long as they require separate pools of resources.

The multiple resource pool theory proposes three dimensions that determine how attentional resources are allocated for concurrent tasks and to which degree concurrent tasks can be performed efficiently:

Information processing modalities: visual vs. auditory. Previous studies suggest that people are better at dividing attention across modalities than dividing it within a single modality. Xu et. al. (2005) presented that users could process and integrate information from visual and auditory channels more efficiently than processing the same amount of information in a single modality.

Information processing stages: perceptual, central processing, and response. Evidence indicates that resources used for an early stage (i.e., perceptual or central processing) are different from resources used for a later stage (i.e., response). This is why a driver can easily monitor the road (perceptual) while steering (responding), but will be more inclined to accidents when steering (responding) and talking on cell phone (responding) (United States National Highway Traffic Safety Administration, 1997).

Information processing codes: imagery/spatial information vs. auditory/verbal information. Research shows that imagery/spatial and auditory/verbal processing requires distinct resources (Wickens & Liu, 1988). This has been agreed and explained by the working memory theory proposed by Baddeley and Hitch (1974 and 2000). The next literature section will explain this in details.

More research has confirmed and expanded the multiple resource pool theory.

Some studies suggest that limited capacity applies to multiple resource pools when people perform time-sharing tasks. For instance, performing concurrent tasks utilizing different modalities sometimes degrades performance on all tasks (Wickens & Ververs, 1998). People are slower in responding to visual stimuli when they are required to monitor the auditory channel at the same time (Spence & Driver, 1997).This is possibly because that the total amount of attention for multiple modalities is limited, and /

or the integration of information from multiple modalities results in a heavy cognitive workload.

Some studies extend the understanding in the factors that determine how well attention is divided among concurrent tasks. Wickens et. al. (1998) proposed four determining factors: (1) the degree to which a task can be performed automatically; (2) the individual's skill in attentional resource allocation; (3) the degree to which information from different modalities is similar which can cause confusion in parallel processing; and (4) the amount of shared attention. With regard to (1), when an automatic processing task is concurrent with another less automatic task, more attention is given to the latter.

2.4.2 Working Memory in Relation to Multimodal Interaction

Working memory, which was previously called short-term memory, operant memory, or provisional memory, refers to the structures and processes used for temporarily storing and manipulating information. Although many working memory models have been proposed, Baddeley and Hitch's model is most commonly accepted and explains why a speech / touch coordinated input mechanism could work well for users.

Baddeley and Hitch's three-component working memory model (1974) potentially explains how integrated speech and hand input can work effectively for human cognition. Their model specifies two subsystems and a central executive. The two subsystems are the phonological loop and the visuo-spatial sketchpad, which are short-term storage and process systems dedicated to distinct content domains.

The phonological loop deals with auditory and verbal process. Visually presented textual information can be transformed into phonological codes by silent articulation before entering the phonological loop. The phonological loop consists of two components: a short-term phonological store that temporarily remembers speech sounds, and an articulatory rehearsal component that repeats the words to prevent them from decay. If a person who is reading is asked to say something irrelevant aloud, his articulatory rehearsal process is blocked which impairs his memory for the verbal material his is reading. This effect is called articulatory suppression.

The visuo-spatial sketchpad holds imagery and spatial information. It consists of a visual cache that stores shapes and colors, and a spatial component that deals with spatial and kinaesthetic information (Logie, 1995). This structure is supported by evidence obtained from previous research, which indicates that there is little interference between visual and spatial tasks, and that brain damage sometimes results in impairment in one component but not the other.

The central executive is an attentional control system that supervises and coordinates cognitive processes in the phonological loop and the visuo-spatial sketchpad if tasks have to be processed in the two subsystems simultaneously. The central executive directs attention to relevant information and inhibits process of irrelevant information.

Baddeley expanded this working memory model in 2000 by adding the third subsystem, *the episodic buffer*, which, Baddeley believes is where information from the other two subsystems is integrated into episodes with chronological order and linked to long-term memory (Baddeley, 2000).

In addition to the structure of working memory, Baddeley specifies that all four components have limited capacity.

Other researchers have argued that the working memory model should be revised to represent a multiple-resource model (Navon and Gopher, 1979; Wichens, 1980 and 1984), because evidence has shown that verbal and spatial working memory are independent to each other (Shah and Miyake, 1996).

This working memory model is supported by experimental results with dual-task diagram and research in brain damages. Dual-task paradigms are used in experimental psychology. They require individuals to perform two tasks simultaneously and compare the performance with single-task conditions. More details can be found in Mousavi et. al. (1995), Woodhead and Baddeley (1981) and Cocchini et. al. (2002).

This working memory model makes two predictions: (1) if two tasks make use of the same subsystems, they cannot be performed successfully together, and (2) if two tasks make use of different subsystems, it should be possible to perform them together as well as separately (Eysenck, 2004).

This working memory model further predicts that in multimodal interaction with integrated speech and hand input, speech tasks and hand movement tasks are processed in two subsystems of the working memory separately, and can potentially be performed simultaneously with little interference against each other.

In general, the above research in cognitive psychology has served as the theoretical foundation for the proposed speech and touch input.

2.5 Summary of Literature Review

In this section relevant literature has been reviewed. Besides providing rich background information, the literature serves as a guideline for designing the exploratory study and the experiment and structuring the research.

2.5.1 Designs for Visually Impaired Users

The literature in the field of designing information access for visually impaired users provides overviews of the fashion and design techniques of the present non-visual information access, as well as revealing the design issues that have not been considered in these products. Specifically, the two major design issues are:

(1) The use of arrow keys in navigation forces users to browse information sequentially. Users need to go through a large amount of irrelevant information to reach the desired information. This type of sequential browsing does not help with the establishment of mental models for comprehending hierarchical information structures.

(2) The prevalent use of keyboards as the main input device requires intensive user learning and memorization. If the functions of keys are not memorized, the functions are not likely to be used unless the user reviews the help document or the user manual to find the function. But relying on the help document or the user manual reduces the usability of a system (Mack et al. 1983; Nielsen, 1994; Nielsen et al. 1986). Furthermore, the user has no "preview" of the command before pressing the key. This means additional effort to reverse actions selected in error.

New interaction mechanisms are needed to address these design issues. The new design should provide both direct access and access through browsing to information and commands. Multimodal input offers potential for providing these interaction alternatives.

Speech input can be designed with a rich command vocabulary and is suitable for directly accessing subjects and instructions. Hand motions on a physical space have proved effective in menu browsing and selection (Friedlander et al. 1998; William and

Tremaine, 2001; Chen et al. 2003, 2004, and 2005). Thus, a new interaction mechanism that integrates speech and hand input is proposed and developed for non-visual information access for the visually impaired.

2.5.2 Designs of Speech and Gesture Interaction

The literature in the fields of gesture input design and speech dialogue design has provided a set of development suggestions for this multimodal speech and hand input mechanism.

When designing a hand gesture input mechanism:

(1) Developers can make use of manipulative gestures, semaphoric gesture, and or pointing gestures in the command grammar. These gestures are broadly used in human communications and human computer interactions. Especially, pointing gestures on a physical space tied with audio or tactile feedbacks have proved effective (Roth et al. 1998 and 2000; Friedlander et al. 1998; William and Tremaine, 2001; Chen et al. 2003, 2004, and 2005).

(2) Using real spatial references as opposed to imaginary objects for gesture input will make locating a desired point or area much easier.

(3) Applying physical constraints as opposed to software feedback will give users more direct guidance on gesture movement and reduce a user's cognitive load in finding physical input locations.

(4) Gesture commands should be implemented on a small working space as opposed to a large working space. This reflects the user patterns observed in both natural gestural tasks and human-computer interactions.

When designing speech input dialogues:

(1) Designers can use finite-based, template-based, or agent assisted dialogues.

(2) Designers can implement a system-initiative, user-initiative, or mixed-initiative dialogue control strategy. Studies have proved that different dialogue control strategies do not necessarily lead to different efficiencies in interaction. User

satisfaction with each of the dialogue types is equivalent. Designers should choose a dialogue control strategy based on the tasks to be performed using the system.

(3) Short speech commands composed of single words or simple phrases achieve a better user acceptance than longer commands because short commands are easier to learn and introduce less system recognition errors. A longer command that compounds several short commands does not necessarily increase the efficiency of the user-system interaction or user satisfaction.

(4) When speech dialogues are designed for a non-visual interface, the system functionality and operation boundaries are invisible. System prompts and verifications are pivotal to guiding user-system interaction. Developers can refer to (Brennan and Hulteen, 1995; Yankelovich et al. 1995; Yankelovich, 1996) for practical guidelines for prompt design.

(5) When speech dialogues are designed for a non-visual interface, the vocabulary of subject and verb commands should be different from the phrases used on a visual interface with similar functionality. The information organization and information flow controls should be redesigned as well. This is because, in a non-visual condition, users' mental models of the information organization and information flow, as well as the vocabulary used to control information display, may be different. Completely mapping a non-visual interface with a visual interface in the same application domain can introduce usability problems.

(6) Speech output is sequential and slow. As such, tradeoffs need to be made between the completeness and the conciseness of the speech feedback. The interaction mechanism should allow users to skip, move forward, and interrupt quickly.

The lessons learned from previous research and design practices have provided valuable guidelines to the design of the AudioBrowser system.

2.5.3 Theories in Cognitive Psychology Related to Multimodal Interaction

The literature in cognitive psychology provides insights as to how human attention and working memory support multitasking in multiple modalities:

(1) Although having limited capacity, human attention can be divided to process multiple concurrent tasks that use resources from different resource pools:

- People are better at dividing attention across modalities (e.g., visual and auditory) than dividing it within a single modality;

- Resources used for an early information processing stage (i.e., perceptual or central processing) are different from resources used for a later information processing stage (i.e., response);

- Spatial imagery/ and auditory / verbal processing require distinct attentional resources.

(2) How well attention is divided among concurrent tasks depends on four factors:

- The degree to which a task can be performed automatically – when an automatic processing task is concurrent with another less automatic task, more attention is given to the later;

- The individual's skill in attentional resource allocation;

- The degree to which information from different modalities is similar (similar information can cause confusion in parallel processing); and

- The amount of shared attention.

(3) Working memory which handles the processing of input from the computer and is used to develop and generate commands to the computer system, consists of four components:

- The phonological loop that stores and processes auditory and language related information;

- The visuo-spatial sketchpad that processes visual, spatial, and possibly movement and kinesthetic information;

- The episodic buffer where information from the other two components is integrated into episodes in chronological order and linked to long-term memory; and

- The central executive that coordinates and monitors the cognitive processes in the phonological loop and the visuo-spatial sketchpad.

This structure of the working memory and the information processing theories from cognitive psychology provide a theoretical background for guiding the design of a multimodal input mechanism that combines speech input and hand gesture input. Users input operations can be executed through speech and gestural channels separately but

concurrently, using different pools of cognitive resources in a separate but integrated manner in the working memory.

2.5.4 Designs of Multimodal Interaction

The literature in the field of multimodal interaction that comprises speech and hand inputs to GUIs has provided insights on users' multimodal behavioral patterns that affect the design of such interfaces. The main findings are:

(1) When given multiple input modalities, users generally use them in a multimodal fashion, but not all tasks are performed using multiple modalities. Map related tasks are mainly performed using integrated speech and pen input rather than speech or pen input only, because pen input is more efficient in drawing and pointing. Speech, when used for the same task, results in speech disfluencies. When using speech recognition systems, direction-oriented tasks are performed using keyboard and mouse more than using speech commands, because use of the keyboard and mouse are more efficient in these tasks, while speech commands are more efficient in target-oriented tasks.

(2) Users use multimodal, rather than unimodal input, to increase their interaction efficiency. Speech utterances during multimodal input are briefer and less complex compared to those in speech-only input. Speech disfluencies are reduced in multimodal input resulting in shorter task completion times.

(3) Each user follows one of the two multimodal integration temporal patterns: simultaneous input or sequential input. The user follows his/her integration pattern most of the time during the multimodal interaction. Changes in user tasks and changes in system recognition error rates do not necessarily result in changes in users' multimodal integration patterns.

(4) When errors occur, users switch modalities to improve the error correction results. Specifically, when one input mode fails, users tend to switch to another input mode to overcome errors.

The behavior patterns observed from previous research have proved users' preferences in multimodal input. However, these behavior patterns cannot be used directly to guide the integration of multiple input modalities in a non-visual information access for visually impaired users, because first, the behavioral patterns are detected on

graphical user interfaces designed for sighted users who may use multimodal systems differently from visually impaired users, and second, the domains of the systems researched are map-related applications and speech recognition applications, which have significant differences from non-visual textual information accessing systems in terms of functionality and user-system interactivity.

Empirical research is needed to test whether user behavioral patterns in multimodal graphical user interfaces can be extended to the proposed non-visual multimodal interaction, and whether the proposed non-visual multimodal input mechanism can be learned and used successfully. The research should also uncover any different behavior patterns of visually impaired users during the interaction. The findings in this empirical research will serve as the theoretical basis for integrating speech and gesture inputs into a non-visual textual information access system for the visually impaired.

CHAPTER 3

RESEARCH QUESTIONS, RESEARCH APPROACH AND AUDIOBROWSER SYSTEM

3.1 Overview

This chapter presents the initial research questions derived from the literature. This is followed by descriptions of the research approach and the AudioBrowser system, which is an eyes-free information browser used for carrying out this research.

The research questions ask that, when performing non-visual information browsing tasks, (1) whether users choose to use multimodal or unimodal input, (2) how they use the multimodal input, (3) how they handle errors in the input, (4) whether the order of training for the modalities affects users' input usage, and (5) whether sighted users and visually impaired users use the input mechanisms the similar ways.

To answer these questions, an exploratory study with sighted subjects was conducted first. The study aimed at refining the research questions and forming testable hypotheses. A controlled experiment was then conducted with visually impaired subjects to test the hypotheses.

3.2 Research Questions

The literature review has indicated that multimodal speech and hand input has great potential to improve visually impaired users' information browsing. The goals of the proposed research are (1) to explore whether a multimodal speech and hand input mechanism will provide solutions to some design issues of existing information access

for visually impaired users, and (2) to explore how users use the multimodal interface and what design indications can be derived from this use.

To achieve these goals, the following research questions have been generated:

RQ1: When interacting with a non-visual multimodal system, do users use multimodal or unimodal input?

RQ2: If users choose to use multimodal input, do they have special multimodal input patterns, i.e., is there a relationship between the type of input operation and users' choice of input modality?

RQ3: What are users' error correction strategies for the non-visual multimodal interface?

RQ4: Does training affect users' multimodal input behavior?

Finally, it will be valuable to ascertain if sighted users use a non-visual system to explore information differently from visually impaired users, because most system designers and developers are sighted and it is therefore difficult for them to understand the difference between their own navigation and visually impaired users' navigation in the information space. The knowledge acquired in this thesis will help to establish guidelines for system developers whose products need to accommodate visually impaired users' needs. Thus, the following research question is proposed:

RQ5: Can we conclude any common or different patterns existing in sighted and visually impaired users' multimodal interaction?

3.3 Research Approach – Exploratory Study and Controlled Experiment

This research combined an exploratory study that determined the parameters of the investigation with a controlled experiment that then tested the hypotheses generated by the exploratory study. The exploratory study was an observation-based gathering of data that would direct the experiment design towards answering the posed research questions. The results from the exploratory study helped to extend research questions, construct testable hypotheses, generate viable stimuli and suggest the controls that needed to be exercised in the experiment.

Due to the limited number of visually impaired participants available, the multimodal interface design which did not give sighted users a significant advantage, and the purpose of the pilot study – to gather information for running a study with visually impaired users, the exploratory study was conducted with sighted users. The controlled experiment was conducted with visually impaired users. Instructions and approaches in the sighted users study were necessarily different than the experiment design in the visually impaired users study, because the study was exploratory and because certain adaptations that were made for visually impaired users would have made the pilot study ponderous, confusing and strange for its subjects, e.g., reading all instructions out loud. This made the pilot study different from the experiment so that data collected in each study could not be compared with statistical tests. However, this arrangement did allow observation-based qualitative comparison of the multimodal interaction patterns between sighted and visually impaired users.

3.4 System Description

The research was conducted on AudioBrowser, a system with both speech and hand inputs. AudioBrowser inherits the hand input on a touchpad from SoundNews (Williams and Tremaine, 2001). The speech input parallel to the hand input was designed and implemented in 2005 as part of this thesis.

AudioBrowser is a non-visual information browser that organizes information, e.g., news articles, into a hierarchy and reads the information for users using auditory output. Users can control the way the content is read using speech and hand input commands.

The hand input is performed on a touchpad, consisting of pointing gestures on the sensing area of the touchpad, and clicks on the buttons beside the sensing area. The speech input is received via a microphone and processed by the Microsoft Speech Recognition Engine. The AudioBrowser system outputs speech and/or non-speech audio displays. Speech output is standard American English synthesized by a Microsoft Text-to-Speech Engine. Non-speech outputs, such as clicks and other sound effects, are prerecorded audio clips.

The designs of the touchpad input and the speech input comply with the design strategies suggested by the literature on gesture input design and speech dialogue design.

The literature suggests that gesture commands should be implemented using real spatial references as opposed to imaginary references. The literature also suggests that gestures should be performed in a small working space that adapts to human natural gesturing. In AudioBrowser, a Synaptics touchpad (Figure 3.1) is used for the touchpad input. It is about half the size of an average human palm. The programming interface

provided by Synaptics allows developers to program functions for the sensing area and the buttons. The sensing area of the touchpad is divided into three tracks. The borders of the tracks are demarcated using paper clips taped to the touchpad. These tracks provide physical constraints through tactile feedback, a feature consistent with suggestions in the literature. The two buttons at the two sides of the sensing area are used to execute commands. The two buttons under the sensing area are not used in the experiment, but have the function of an abort button for AudioBrowser.

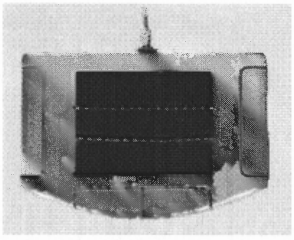


Figure 3. 1 Programmed Synaptics Touchpad

The tracks in the sensing area are dynamically divided into small segments, as illustrated in Figure 3.2. One information item or operation command is mapped to each segment. When the user's finger touches a segment, the system speaks aloud the corresponding item on the segment. When the user reaches the boundary between two segments, the system outputs a "click" sound to indicate that a boundary is being traversed. This allows a user to rapidly traverse 3 or 4 segments without listening to the underlying speech contained in the segment. When the user proceeds to the next segment, the system halts the speech output of the previous segment and then outputs speech for the new segment being pointed to.

The first track is dedicated to browsing information that is organized hierarchically (Figure 3.2). The information is dynamically divided into segments that

map onto items on the information level being navigated. A user can explore the items at that information level by gliding a finger across the track. If an item being pointed to is a category, pressing the right button selects the category and goes down one information level. The same track is then rewritten with information items at the new level. Pressing the left button goes up one level in the information hierarchy.

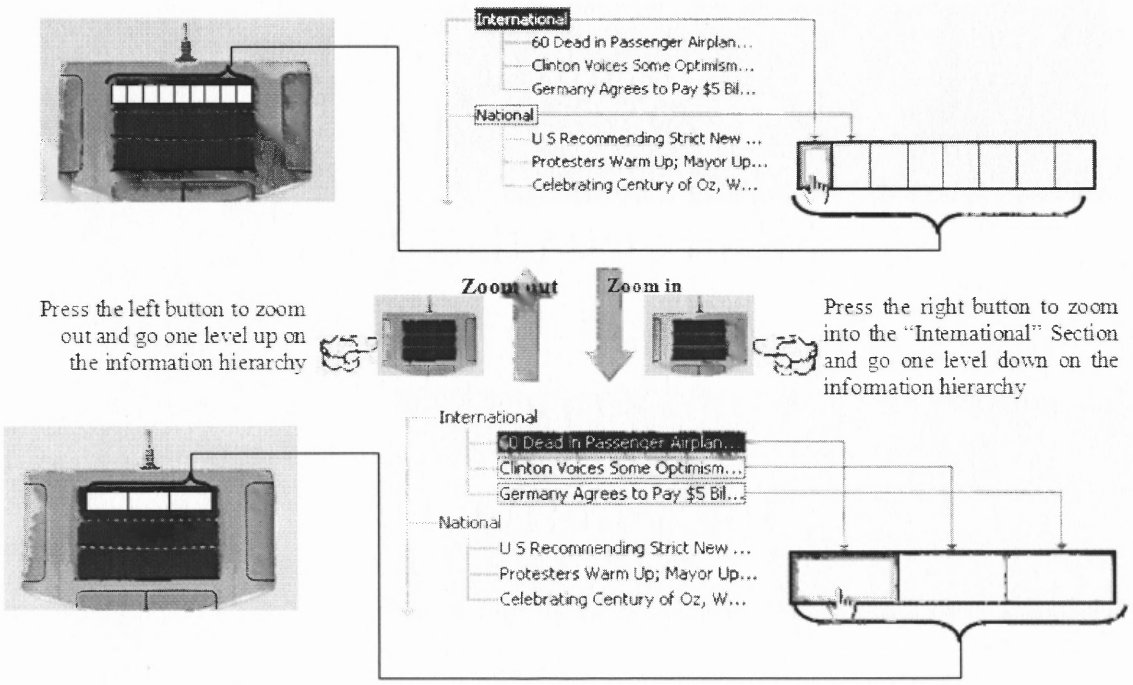


Figure 3. 2 Browsing Hierarchical Information Using the Touchpad

The second and third tracks are mapped with operation commands. Users explore these tracks to find a command and then press one of the buttons to execute the command. Button clicks are mode-sensitive, i.e., what function a button click performs depends on what item is selected in the track. For example, when the command "read by sentence" is selected, clicking on the right button reads the next sentence, and clicking on the left button reads the previous sentence. When "change reading volume" is selected, clicking on the right button increases the reading volume and clicking on the left button

decreases the volume. This is a solution to deploy functions on a limited operation space, but not to lose intuitivity and consistency of the operations, as clicking the right button always results in an action going to "next" or increasing a value, and clicking the left button always results in an action going to "previous" or decreasing a value.

The speech input grammar size is fifty-four commands. Each command consists of one to four words, while the majority of commands consist of two or three words. This simple and fixed grammar design is consistent with the literature's design suggestion that isolated word recognition for speech input in a fixed domain with limited tasks is sufficiently accurate for effective user interaction.

Speech dialogues are user-initiated. A speech command will be executed when the user holds down a "push-to-talk" button (i.e., the ctrl button on the keyboard as programmed) and speaks a command into the microphone.

Rather than using complicated compound speech commands such as "Find the email from John Smith on May 1st 2005", simple and short speech commands that parallel the touchpad commands were created for the speech grammar. This means that any input operator (i.e., the smallest input operation unit) of a user task can be done using either speech input or touchpad input. Performance of a user task usually comprises a series of input operators. Thus, a user task can be performed using either multimodal input or unimodal input, depending on a user's choice.

There are three reasons for creating parallel speech and touchpad commands at the operator level. First, the user can freely decide to finish a user task using mixed input modalities. Second, at any time, when one input mode fails, the user can switch to the other to recover from the failure. Third, a speech command that is forgotten can always

be found on the touchpad. The most frequently used touchpad and speech commands are listed in Table 3.1.

Table 3. 1 Frequently Used Touchpad and Speech Commands

Typical User Task	Touchpad Commands	Speech Commands
Browse information on one level in the information hierarchy	Glide across the first track	"Next article/ category/ item", "Previous article/ category/ item"
Go to a different information level by entering or exiting an information category	Click the right button to enter an information category, click the left button to exit	"Select" or "zoom in", "exit" or "zoom out"
Set the text unit (i.e., word, sentence, paragraph, or complete content) by which the system reads the content	Find the command "set to word", "set to sentence", "set to paragraph", or "set to complete content" on the second track	"Set to word", "set to sentence", "set to paragraph", "set to complete article"
Read the next or the previous text unit	Find "set to word", "sentence", or "paragraph" on the second track, click the right button to read the next unit, or click the left button to read the previous unit	Method 1: "Set to word/ sentence/ paragraph" + "Next" or "previous" Method 2: "Next word/ sentence/ paragraph" or "Previous word/ sentence/ paragraph"
Browse the text unit commands, in case the user forgets the text units available for use	Glide across the second track	"Output unit" (by which the system speaks aloud all the text units available for use)
Pause reading and resume reading	Click the right and the left buttons together to pause, to resume click the right button (which leads to reading the next text unit) or click the left button (which leads to reading the previous text unit)	"Pause", "resume"
Spell a word	Find the command "set to word" on the second track and click the left and the right buttons together to spell the current word	"Spell" or "spell word"
Browse system settings, in case the user forgets what settings are available for adjusting	Glide across the third track	Method 1: "Settings menu" (by which the system speaks all the system settings) Method 2: "First setting" + "next setting" or "previous setting" (by which the system speaks one setting at a time)
Change reading speed/ volume/ pitch	Find "speed", "volume" or "pitch" on the third track and click the right button to increase it, or click the left button to decrease it	Method 1: "increase speed/ volume/ pitch" or "decrease speed/ volume/ pitch" Method 2: Speak "First setting" and repeat "next setting" until the wanted setting is found, then speak "increase" or "decrease"
Change reading voice	Find "voice" on the third track and click the right button to use the next voice on a voice list, or click the left button to use the previous voice on the list	Method 1: "next voice" or "previous voice" Method 2: Speak "First setting" and repeat "next setting" until "voice" is found, then speak "next" or "previous"

3.5 Design Issues Encountered and Solutions Implemented During Iterative System Development

This section describes the design issues encountered and their solutions during the iterative system design and implementation process of building a multimodal AudioBrowser.

First, the speech recognition engine is designed to process everything it hears, even irrelevant sounds. This characteristic caused problems and a method was needed for disambiguating speech commands from irrelevant sounds. The first implementation of AudioBrowser with the speech recognition feature emphasized the severity of this problem. The system "heard" irrelevant speech and noise, interpreted them as speech commands, and executed the closest-sounding command on a constant basis even when the user was intending to only use the touchpad. For example, a door closing in the next laboratory easily led to the interpretation of the "next" command.

A microphone with an on/off button was tried as a way of fixing this problem. The user switched the microphone on to speak a command and switched it off when finished. Unfortunately, turning the microphone on and off generated an interfering sound that was also interpreted as speech. In addition, there was latency between the button being switched on and the system starting to process sounds so that the spoken command was not heard but ambient noise was processed leading again to an unexpected command being selected.

A "push-to-talk" button on the keyboard of the computer was then tried. This eventually became a feature of the system. The button needed to be pushed down in order for the system to execute a spoken command. This approach filtered out many irrelevant

sounds, and the button did not generate interference with the speech recognition. An inconvenience is that users still need to pay attention to a small latency and push the button before, not during their utterance of a command. The "Ctrl" button on the keyboard is used to fulfill this "push-to-talk" function.

One problem not solved by the first "push-to-talk" idea was that of the button irrelevant sounds still being accepted by the system when the button was pushed. Because of this, one push could result in the system's recognition and execution of a mixed series of correct and incorrect commands. In many circumstances the system's speech output interfered with its recognition of user speech commands, and formed an execution loop. To significantly reduce this symptom, each time the button is pushed, the system recognizes only the first heard command. As long as the user does not wait a long time before speaking after pushing the button, most irrelevant sounds are filtered out successfully.

The second issue was that some speech commands were harder to recognize by the recognition engine than other commands. Changes were made in the speech grammar to avoid these recognition problems. AudioBrowser understands a fixed list of speech commands. A command can be a single word or a phrase. Shorter commands take a shorter time to be spoken, but risk more speech recognition errors. Longer commands, on the contrary, require only a little more overheard in time, but reduce the risk of speech recognition errors. The speech grammar has been gradually adapted to reach a balance between the length of the commands, the naturalness of the language, and the risk of recognition errors.

An example of this balance can be seen in the original speech command of "word." "Word" had been a speech command initially for setting the system to read text word-by-word. Due to its short length and high phonetic similarity to other sounds, the system constantly recognized irrelevant sounds in the environment as the command "word". After the command "word" was changed to "set to word" the recognition error was significantly reduced.

This chapter serves as a prelude to the next chapter. Having presented the research questions of interest and the system that will be used to test out the research questions, an exploratory study is needed to refine these research questions and test the feasibility of the research being underdtaken. The next chapter presents this exploratory study.

CHAPTER 4

DESIGN OF EXPLORATORY STUDY WITH SIGHTED USERS

4.1 Overview

The goal of the exploratory study was to determine if the patterns in multimodal integration described in the research questions, could be observed. A speech input grammar was designed for AudioBrowser that paralleled the touch input grammar that already existed in AudioBrowser. This parallelism had the advantage that one modality could replace another modality at any time in the interaction. This potential replacement meant that the study could also examine if multimodal interfaces were effective for helping with error correction, that is, if one modality caused an error, a second modality could possibly be selected and the input command redone, thereby avoiding the error.

The entire study was set up to be as naturalistic a study as possible, i.e., both types of input (speech and touch) were taught to the subjects, but choice of which modality to use for any operation was completely at the subject's discretion. The subjects in the study were asked to perform tasks that they would normally do with AudioBrowser. As such, the study was more of an exploratory study than a controlled experiment except that all subjects received the same training and the same tasks to perform with AudioBrowser. The study sessions were captured on video. The subjects were also invited to provide comments and opinions in a post-study interview.

The next section of this chapter describes the subjects who participated in the study, the procedures carried out in the study, the tasks given to the subjects, and the data

that was captured. This is followed by a section describing the coding done on the qualitative data collected.

4.2 Subjects, Procedure, Tasks and Data Capture

Fifteen subjects participated in the exploratory pilot study. During the study it was found that one subject did not speak English fluently enough to use the speech recognizer in AudioBrowser. Hence, his data was excluded from this report because of the large number of speech understanding errors that occurred. The remaining fourteen subjects all spoke fluent English. At the time of the study, four subjects held a masters degree or were enrolled in a Ph.D. program and ten were undergraduates. The graduate level subjects participated in the study as volunteers. The undergraduate students participated for course credit, as part of an information systems evaluation course. The students were given the options of performing a system evaluation project individually, or participating in the exploratory study to help evaluate AudioBrowser. All subjects majored in computer science, information systems, information technology, or human computer interaction.

The subjects were sighted but were not provided any visual interface during the study. The subjects interacted with the system by listening to the auditory output. During the study the subjects sat in a laboratory room with a preprogrammed Synaptics touchpad, a directional microphone, and a regular keyboard in front of them. A directional microphone receives sounds from a specific direction and restricts sounds coming from other directions. The "Ctrl" button on the keyboard was used as the "push-to-talk" button for speech input. The equipment was connected to the AudioBrowser system. The user's speech input was received and processed by the Microsoft Speech

Recognition Engine. The system's speech output was standard American English synthesized by the Microsoft Text-to-Speech Engine. The non-speech audio output such as the "click" and the "whish" sounds used to indicate a change of state in AudioBrowser were prerecorded sound clips.

The complete study for each subject took three consecutive days. The entire participating time of each subject was approximately six hours. In the first two days, the subjects participated in two tutorials on how to perform speech input and touchpad input. Preceding the speech input tutorial, each subject spent approximately thirty minutes in training the speech recognition engine to their speech patterns. Half of the subjects had the speech input tutorial in the first day and the touchpad input tutorial the second day. The other half received the tutorials in the reverse order. The subjects were assigned to the two training orders randomly. During the tutorials, the subjects read written instructions and tried the system functions as directed by the tutorial document. A practice session followed each tutorial session. In the practice session, the subjects performed a list of tasks that covered all the system functions using the input method taught in the preceding tutorial session.

On the second day when the subjects had learned both speech and touchpad input, they were asked to freely mix the two input methods to do a new set of tasks. This allowed them to find their own multimodal strategy. The tutorials and practices in the first two days were ample so that each and every subject mastered the input methods and formed his/her multimodal pattern of input.

On the third day, the subjects first warmed up by using their selection of multimodal inputs to finish a set of tasks. Once this warm-up was complete, the study

session began. Subjects were given a new list of tasks that represented the complete system functions and the typical use of the system. The tasks included browsing information categories and finding a particular piece of information, reading the information by specified text units (e.g., paragraph by paragraph), proceeding forward or going backward within the text to find a particular name to spell, comprehending a paragraph, pausing and resuming reading, changing the reading speed, volume, and so on. These tasks covered all system features and represent a real information-browsing scenario using the system. The subjects were not restricted as to what input methods to use and how to use them. A digital video camera fixed on a tripod was used to capture all user input operations and system responses.

Table 4. 1 Exploratory Study Procedure

	Day	Subject Group 1	Subject Group 2
Learning and individual multimodal patterns formation	Day 1	Study introduction Consent form filling out Background questionnaire filling out	Study introduction Consent form filling out Background questionnaire filling out
		Speech recognition engine training Speech input tutorial Speech input practice	Touchpad input tutorial Touchpad input practice Speech recognition engine training
	Day 2	Touchpad input tutorial Touchpad input practice	Speech input tutorial Speech input practice
		Speech input review	Touchpad input review
		Multimodal input practice	Multimodal input practice
	Day 3	Multimodal input warming up	Multimodal input warming up
		Experiment	Experiment
		Post-study interview and questionnaire administration	Post-study interview and questionnaire administration

After the experiment session, the subjects were interviewed about their experience with AudioBrowser. The interview was conducted as follows: The experimenter played back the video of the experiment session task by task. The subject watched the video and

described what he or she was doing, why a particular input method was used to perform the task, what problems had been encountered, and any comments he or she had about the input method chosen. Subjects were then asked to fill out a post-study questionnaire.

4.3 Experimenter Notes Preparation and Coding

The experimenter transformed the videos of each subject's experiment session into comprehensive notes documenting the corpus of user inputs and system responses. Notes were taken based on operators, the smallest user input unit. For each subject, the documentation included operator series by the subject to finish an experiment task, input mode used operator by operator, whether an operator succeeded or not, the problems that happened, the remedy the subject took, and the subject's comments. Coding was then conducted based on the notes for the following variables:

Input mode use: Input modes were coded for each operator. They are either speech input or touch input.

Input mode switches and potential switch causes: Input mode switches were marked out. Based on the experimenter's observation, there are four potential reasons for input modality switching. (1) *Operation Type Change*: when the operation type changed between navigation and non-navigation operations, the subjects changed input mode. (2) *Mode Failure*: when failures in user input occurred, the subjects switched to a different input mode to recover. (3) *Experiment Task Change*: when one experiment task was finished, a subject stopped using the system to read the next experiment task. This hiatus may have reset the subject's choice of input mode. (4) *Need of repetition*: the subjects

seemingly preferred the touchpad input for repetitive operations. When they needed to perform an operation repeatedly, they switched to the touchpad to execute this.

Operation types: Operation types reflect the goals of input operators. For example, to finish the experiment task "find a news article about Hillary Clinton", the user needs to use two types of operations: "browse the information categories" and "zoom into a category".

The operation types involved in using the AudioBrowser system are in two application domains: information browsing/reading and system settings control. The detailed operation types in the two domains are illustrated in Table 4.2.

Table 4. 2 Operation Types

Domains	**Operation Types**	**Details**
Information browsing/ reading	Browse a single level information	Go to the next or previous article or information category
	Go to a different information level	Enter or exit an information category
	Proceed or recede within a text	Read the next or previous word, sentence, or paragraph
	Non-proceeding or -receding operations within the text	Pause
		Resume
		Spell a word
		Read the current article again from the beginning
		Repeat
	Set reading unit	Set to word, sentence, paragraph, or complete content
	Show the list of reading units	
System settings control	Search for a system setting	Search for reading speed, volume, voice, pitch, and the speed of non-speech sounds
	Change the value of a system setting	Increase or decrease speed, volume, or pitch; use the next or previous reading voice
	Show the list of system settings	

The detailed operation types can be abstracted to two higher level types: *navigation operations*, which relate to searching and locating an item in the information space or the command space, and *subject or verb instructions*, which are abstract commands not tied with locations in the information and the commands spaces.

***Repetitive operations*:** When the user gave a command exactly the same as the immediately preceding command, the user's operation was a repetitive operation. Repetitive operations were recorded for two causes: (1) The user repeated a command because the preceding command failed (i.e., the preceding command did not achieve the user's goal either because the user used a wrong command or because the system mis-recognized the command); (2) The user repeated a command to achieve a system response the same as the preceding response.

***Successful, partially successful, and failed operations*:** Every operator was recorded for its success, partial success or complete failure. Successful operators are those that resulted in system responses that users aimed at. Partially successful operators represent the situations where the system reacted to the user's input operation with correct system executions, but either that the system did not deliver the auditory feedback to the user clearly, or that the user expected a different result because the user's mental model was different from the way that the system actually worked. Completely failed operations are those led to an explicit error message or those without any system responses.

***Failure types*:** For the completely failed inputs, several major failure types were recognized. The failure types occurred in speech input and in touchpad input were not completely the same.

For speech input, there were six major types of failures. (1) *Speech output interference*: When the speech recognition engine inadvertently recognized its own speech output as a speech command and executed a corresponding command. A typical scenario of this type of failure is that the user said "Pause" when the system was reading an article aloud, and the system was interfered by its reading and recognized the command incorrectly. (2) *Speech recognition errors*: The speech recognition software failed to recognize a clearly uttered speech command when no obvious external interference existed. (3) *No reaction symptom*: A speech command was given by a user, but the system did not respond. (4) *Environment noise interference*: When the speech recognition engine inadvertently recognized environment noise as executable commands. Interference examples include when the push-to-button was pressed, the microphone picked up environment noise, such as irrelevant talk by the subject, the spin of the computer CPU fans, and the echoes reflected by the wall of the experiment room. (5) *Incorrect user mental model*: The user understood the system in a way different from how the system actually worked. (6) *Other failures*: In some occasions the failure was not any of the aforementioned types, such as system bugs.

For touchpad input, there were three major types of failures. (1) *Mode errors*: These represent the major touchpad failure type. The three tracks of the touchpad are basically three modes of the touchpad input. The functions of the two physical buttons aside are based on the present mode of the touchpad, i.e., when it is in the information category/ title browsing mode (when the first track is in use), the two buttons execute zooming in and zooming out functions; when it is in the text unit setting mode (when the second track is in use), the two buttons execute reading the next text unit and reading the

previous text unit functions; when it is in the system setting adjustment mode (when the third track is in use), the two buttons execute increasing the setting value and decreasing the setting value functions. This design was a solution to deploying a necessary set of functions within the limited design space of the touchpad. But this design introduced user operation errors. For example, the touchpad's static nature caused users to forget the present system mode and when a tactile command was executed, an unexpected system response occurred. (2) *Incorrect user mental model*: Similar to the speech input, some operation errors occurred in touchpad input were caused by that the mental model of the user mismatched the way that the system actually worked. (3) *Other failures*: Other failures were mainly caused by system bugs existed.

***Error correction operations*:** When an input error occurred, the user took actions to recover from the error. The error correction operations were coded at the operator level into four categories: (1) an unsuccessful speech operator is followed by a speech operator for error correction, (2) an unsuccessful speech operator is followed by a touchpad operator to fix the problem, (3) an unsuccessful touchpad operator is followed by a touchpad operator; and (4) an unsuccessful touchpad operator is followed by a speech operator.

Error correction cases were also marked. An error correction case contains a sequence of error correction actions until the error is corrected or given up.

4.4 Reliability

The notes and coding were finished by one experimenter. The coding was conducted twice and over 90% of the coding was the same.

CHAPTER 5

RESULTS AND DISCUSSION OF EXPLORATORY STUDY

5.1 Overview

The purpose of the exploratory study was to explore possible causes and relationships that could be tested in a controlled experiment. Abundant qualitative and quantitative data was analyzed. Rich results were obtained.

The most important observations are as follows. (1) All the subjects chose to use multimodal input, while multimodal and unimodal inputs were both available to them. (2) The subjects seemed to choose modalities based on the type of operations they were performing. The subjects used touch input for navigation, and speech for short tasks that interrupted navigation. (3) When error occurred, the subjects stayed in the failing modality instead of switching to another modality to correct the errors.

In addition, the exploratory study found that some factors, which were not included in the original research questions, seemed to affect the subjects' input modality choices and switching. The factors are: cognitive load, error rate, and the level of vision available for task performance.

Based on the observations obtained from the exploratory study, the research questions were revised. Hypotheses based on the observations obtained from the exploratory study were formed for statistical verification. The next chapter provides more details on the revision in the research questions and the hypotheses.

5.2 RQ1: Multimodal or Unimodal

Research question 1 is: When interacting with a non-visual multimodal system, do users use multimodal or unimodal input? To study this question, the number of speech inputs and the number of touchpad inputs were counted. The switches that occurred between input modes were then tabulated and analyzed for possible underlying reasons for each switch.

There were a total number of 1642 input operations performed by the 14 subjects during the experiment sessions, among which 635 or 38.67% were speech input operations, and 1007 or 61.33% were touchpad input operations. The total number of input operations by individual subjects averaged 117.3, ranging from 73 to 200, with a standard deviation of 29.43. The number of speech input operations averaged 45.4, ranging from 5 to 81, having a standard deviation of 24.9. The number of touchpad inputs averaged 71.9, ranging from 27 to 129, with a standard deviation of 30.8. Thus, there was a wide range of individual behavior in modality choice, but all subjects mixed speech and touchpad inputs during the experiment sessions.

5.2.1 Input Modality Switch Analysis

During the total number of 1642 input operations by all subjects, 222 input mode switches occurred, which ranged from 7 to 27 switches for each subject. Through analyzing the experiment session videos, four major potential causes of input mode switches were identified:

Change of Operation Type: As explained, a user task can contain several operators, each of which can be classified into one of the operation types described in

"Operation Types" in the "*Experimenter Notes Preparation and Coding*" section. The change of operation types was a major potential reason of the input mode switch. A total number of 148 or 66.67% input mode switches were accompanied by operation type changes. Among these 148 input mode switches, 133 (i.e., 59.91% of 222) were accompanied only by operation type changes, and 15 were accompanied also by other identified causes of input mode switches.

Operation Repetition: It has been found that when there was a need to repeat a system action, the subject tended to switch to touchpad input to perform the repetition. The need of repetition constituted to 9.91% (i.e., 22 out of 222) of the input mode switches. One of the 22 switched was also accompanied by another potential reason. 21 mode switches (i.e., 9.46% of 222) were performed when only the need of repetition is presented.

Preceding Input Failure: Although the experimenter has observed that the failure of an input operation did not necessarily lead to a switch of the input mode, there were still 36 switches (i.e., 16.22% of 222) that involved input failures in the immediately preceding operation step. Among the 36 occurrences, 6 coinstantaneously involved another potential switch cause, and 30 (i.e., 13.51% of 222) involved only input failures in the preceding step.

Start of New Tasks: When one experiment task was finished, the subject spent some time reading the next experiment task. It was suspected that some switches of input modes were due to this intervention, i.e., after a break in an operation flow, the subject had an opportunity to choose another input mode. 22 (i.e., 9.91% of 222) input switches occurred when the subjects finished one experiment task and started another. 10 of the 22

occurred coinstantaneously with another potential switch cause. The rest 12 (i.e., 5.41% of 222) occurred merely with experiment task interference.

These statistics covered 95.50% input modes switches. The rest 4.50% (i.e., 10 of 222 switches) did not involve any of the four causes above. The following pie chart illustrates these results.

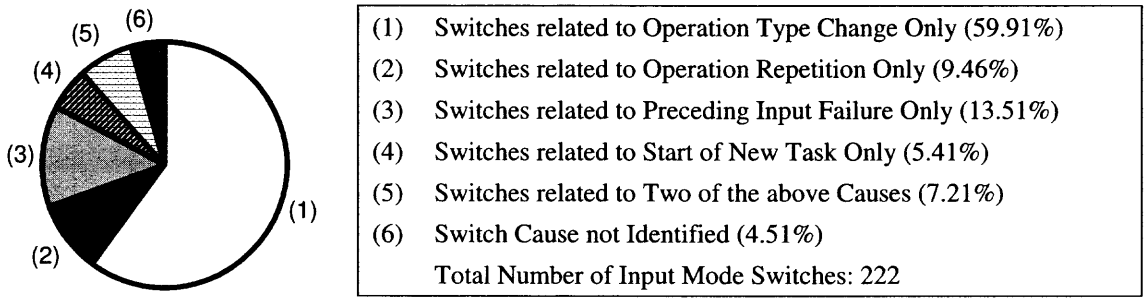


Figure 5. 1 Input Mode Switches Illustrated by Potential Causes

One particular interest of the experimenter's was to investigate whether any of the potential causes of input mode switches were also the predictors of input mode switches. For this purpose, a correlation analysis was conducted between the dependent variable, the total number of input mode switches, and the potential predictors: (1) the total numbers of operation type changes, (2) the total numbers of input failures, and (3) the total numbers of operations to repeat a system action. Since all subjects experienced the same number of experiment task interventions, it is not possible to investigate the correlation between the number of experiment tasks and the number of input mode switches at this time.

Normality was checked within each corresponding data set. For the total number of modality switches and the total number of repetitive operators, identical values existed in the data sets, and hence the Kolmogorov-Smirnov test was adopted. Since no identical values existed in the data sets, it was appropriate to use the Shapiro-Wilks test for the total number of operation type changes and the total number of input failures. The test results indicated that all data sets satisfied the normal distribution assumption. Therefore, Pearson's r is the correct correlation calculation for the data.

Table 5. 1 Test of Normality

	Kolmogorov-Smirnov [a]			**Shapiro-Wilks**		
	Statistic	df	Sig.	Statistic	df	Sig.
No. of modality switches	0.11887292	14	0.2*	0.906902534	14	0.142075
No. of operation type transition	0.176529652	14	0.2*	0.959148734	14	0.709048
No. of repetitive operators	0.225987846	14	0.050901624	0.79642695	14	0.00451
No. of input failures	0.166722319	14	0.2*	0.957557351	14	0.682738

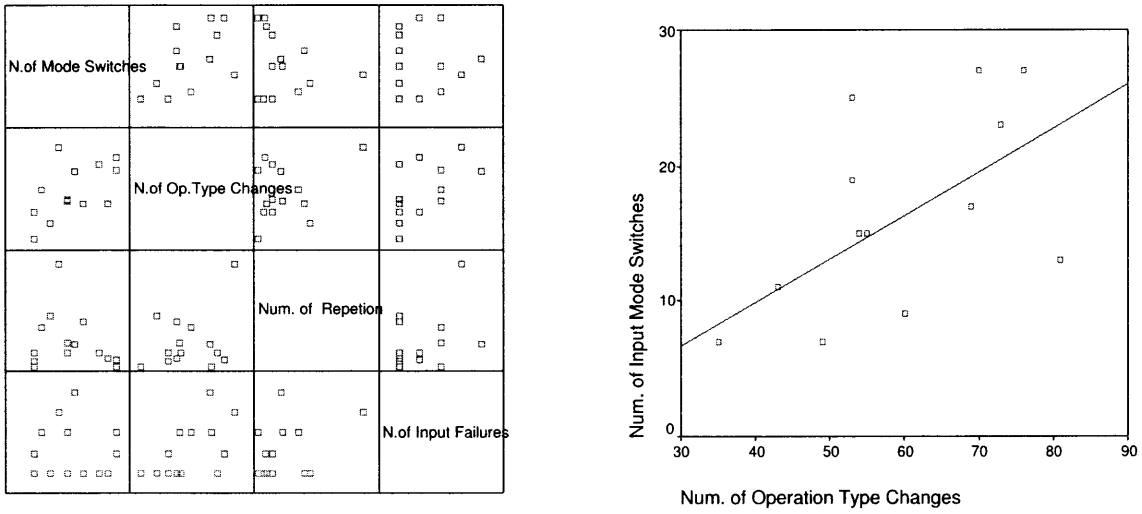
* This is a lower bound of the true significance.

a Lilliefors Significance Correction

It is shown in Figure 5.2 that the best linear relationship exists between the number of task type changes and the occurrences of input mode switches, whose Pearson's r = 0.587 with p (one-tailed) = 0.014.

To further investigate the predictive ability of these factors on input mode switch, a linear regression was conducted. The resulting model with all three factors has R = .773 and R Square = .597. The significance values and the 95% confidence intervals indicate that two of the factors, system action repetitions and input failures, should be dropped from the model. So task type change remains in the model. However, this regression

model with only one predictor is not viable because its predicting ability is not satisfied (with R = .587 and R Square = .344). Nevertheless, the correlation is high enough to suggest that task type change might be a reason for modality switching and therefore, this is kept as a hypothesis to test in the thesis experiment.



Left figure: The figure presents a correlation matrix between the dependent variable (i.e. the num. of input mode switches) and three independent variables (i.e. the num. of operation type changes, the num. of operations to repeat a system action, and the num. of input failures). The purpose of the correlation check is to see whether the independent variables have any predictive effect on the dependent variable. Each red dot in the correlation matrix represents the data of one subject. In each grid there are fourteen dots. The only dependent variable that presents a linear correlation is operation type changes.

Right figure: The figure presents the linear regression based on between the number of input mode switches and the number of operation type changes. Each red dot represents the data of one subject. There were thirteen dots in the figure because two subjects' data superposed at the (49, 7) point.

Figure 5. 2 Correlations between Input Mode Switches and Possible Predicting Factors

Table 5. 2 Coefficients of Linear Regression Models Predicting Input Mode Switches

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-4.898	7.669		-.639	.537	-21.986	12.189
	Task Type Changes	.495	.137	.898	3.627	.005*	.191	.800
	System Action Repetitions	-.171	.099	-.374	-1.722	.116	-.392	.050
	Input Failures	-1.902	1.409	-.342	-1.350	.207	-5.041	1.237
2	(Constant)	-3.098	7.732		-.401	.696	-19.944	13.749
	Task Type Changes	.324	.129	.587	2.510	.027*	.043	.604
Dependent Variable: Input Mode Switches								

5.3 RQ2: Multimodal Input Usage: Input Modality – Operation Type Dependence

Research Question 2 is "If users choose to use multimodal input, do they have special multimodal input patterns – i.e., is there a relationship between the type of input operation undertaken and a user's choice of input modality?" To study this question, the subjects' actual use of input modes corresponding to input operator types was analyzed. The subjects' subjective ratings of the speech input and the touchpad input respectively for each input operation type were also analyzed.

Paired t-test was the general method used for analyses. Since Paired t-test is robust to non-normally distributed data, test of data normality was not necessary.

5.3.1 Users' Choice of Input Modality for Each Operation Type

It was found that when the subjects used AudioBrowser to browse information, their use of input mode for each operator was operation-type dependent. It means that the subjects chose a specific input method for a specific operation type. The results are elaborated

along the nine categories of operation types listed in Tables 5.3 and 5.4. These operation types are at the operator level.

Operation Type 1: Browse single level information: This type involves browsing the titles of the next or the previous information category or article. The operation using the touchpad is gliding the finger across the virtual segments on the information-browsing track (the first track). The speech input operations can be speaking "next category", "previous category", "next article", or "previous article", etc. The average number of touchpad input used by the subjects to perform this type of operation was confirmed to be significantly more than the average number of speech input used to perform this operation type (i.e., 6.29 versus 2.14 input operations respectively, paired t (13) = 3.342, p = 0.0027, one-tailed).

Operation Type 2: Go to a different information level: This type involves entering (i.e., zooming into) or exiting (i.e., zooming out of) an information category. The touchpad operation is to click the side buttons. The speech operation is to say "zoom in" or "select", and "zoom out" or "exit". The average number of speech inputs used to perform this task was 2.64, while the average number of touchpad input used was 5.07. The touchpad operations used were significantly more than the speech operations, with paired t (13) = 1.913, p =0.039, one-tailed

Operation Type 3: Proceed or recede within a text: This type of operation read the next or previous word, sentence, or paragraph. The touchpad operation involves setting the text unit (i.e., word, sentence, or paragraph) which the system should read, and clicking the buttons to go to the next or the previous text unit. The speech operation is to say "next word/ sentence/ paragraph", or "previous word/ sentence/ paragraph". The

results showed that the subjects overwhelmingly preferred to use touchpad than speech. The average number of touchpad input used was 41.07, while the average number of speech used was 9.43. Paired t (13) = 4.569, p = 0.0003, one-tailed.

Operation Type 4: Non-proceeding or -receding operations within a text: Besides proceeding and receding, there is one set of operations to control the information reading. Examples of frequently performed controls are pause, resume, spell word, read the current article from the beginning, and repeat. A comparison between the number of speech commands used and the number of touchpad input used by individual subjects showed that speech was significantly used more than touch (i.e., 19.57 versus 7.21 respectively, paired t (13) = 4.314, p = 0.0004, one-tailed). To show more details, the analysis for individual frequently used command was conducted.

Command 1: Pause: To pause reading, the subjects used slightly more speech input than touchpad input, (i.e., 5.36 versus 3.93, on average). This did not lead to any significant difference in the amount of speech and touch used.

Command 2: Resume: The subjects used significantly more speech input than touchpad input to resume paused reading. The average number of speech and touchpad inputs was 8.79 versus 0.5. Paired t (13) = 7.173, p =0.000004, one-tailed.

Command 3: Spell: No significant difference has been found between the amount of speech input used and the amount of touchpad input used to spell a word. The average numbers are 1.07 speech inputs and 0.93 touchpad inputs.

Command 4: Read the current article from the beginning: Again, no significant difference has been found. The subjects used 1.93 speech inputs and 1.79 touchpad inputs, on average, to perform this task.

Command 5: Repeat: The subjects used significantly more speech than touch to perform this task. The averages were 1.43 versus 0 – actually no touchpad input was used. Paired t (13) = 2.589, p = 0.0112, one-tailed.

Operation Type 5: Set reading unit: This type of operation is to set the reading unit to word, sentence, paragraph, or complete content/article. Once it is set the system reads one unit at a time. When performing it, the operation on the touchpad is to glide the finger across the virtual segments on the second track (i.e., the text unit track) until the segment mapped with the desired text unit is reached. This action on the touchpad is actually also a process of looking for a text unit by using recognition memory. Speech input, on the contrary, requires the user to name the desired text unit directly (i.e., using recall memory) by speaking "set to word", "set to sentence", "set to paragraph", or "set to complete article". Although paired t test did not show any significant difference between the number of speech commands used and the number of touchpad input used, touchpad was used more than speech input (i.e., 7.29 versus 5.57 respectively).

Operation Type 6: Search for a system setting: When users do not remember the system settings available for adjusting (adjusting what?), they need to browse the setting options and use their recognition memory to pick the desired setting. Similar to the touchpad operation for task type 5, the touchpad input for searching for a system setting is to glide across the virtual segments on the third track (i.e., the system settings track), listen to the speech-announced segments, and pick the desired setting. The speech

operations include three speech commands, "first setting" which lets the system point to the first setting on the system settings list, and "next setting" and "previous setting" that move the system pointer to the next or previous setting on the list. When searching for a setting, the subjects used none speech input and 1.64 touchpad input, on average. This contributed to a paired $t(13) = 3.846$ and one-tailed $p = 0.001$.

Operation Type 7: Change the value of a system setting: To increase or decrease the value of a system setting (i.e., speed, volume, voice, pitch, or the volume of the non-speech audio) using the touchpad, the user needs to search for the setting first on the system settings track, then click the buttons to change its value. To perform this operation using speech, the user needs to speak "increase" or "decrease" and the name of the setting, e.g., "increase volume". During the experiment, the subjects used nearly the same amount of speech and touchpad input (i.e., 3.86 versus 3.21, on average). The difference is not significant.

There are two other types of operations in Table 5.1 that were not used by the subjects at all. Those are to show the complete list of text units available for set, and the complete list of system settings available for adjustment. When the subjects could not recall a setting or a text unit, they simply glided through the touchpad tracks to find it, rather than letting the system to read the list of options to remind them.

The findings about the dependency of input mode choice on operation types are summarized in Table 5.3. Figure 5.3 shows the detailed comparison among the operation types.

These operation types, through further analysis, can be classified to two major categories: *navigation operations* and *subject or verb instructions*. Navigation operations involve navigating the information or commands hierarchy to locate an item. They relate to locations either on the information hierarchy or within a node of the hierarchy. Subjective or verb instructions are commands not tied to locations in the information space. The analysis reveals that the choices of input methods for the two categories of operations are significantly different (See Table 5.4). The conclusions are that when performing navigation operations, the subjects used significantly more touchpad input than speech input, while when performing subjective or verb instructions, the subjects used significantly more speech input than touchpad input.

Table 5. 3 Input Mode Choice by Task Types

Operation Type		Avg. Speech Used	Avg. Touch Used	Paired T Test Result (one-tailed)	Observations
Browse a single level information		2.14	6.29	p = 0.0027	14
Go to a different information level		2.64	5.07	p =0.039	14
Proceed or recede within text		9.43	41.07	p = 0.0003	14
Non-proceeding or -receding info-browsing operations within the text	Overall	19.57	7.21	p = 0.0004	14
	Pause	5.36	3.93	--	14
	Resume	8.79	0.5	p =0.000004	14
	Spell	1.07	0.93	--	14
	Read the current article from the beginning	1.93	1.79	--	14
	Repeat	1.43	0	p = 0.0112	14
Set text unit		5.57	7.29	--	14
Search for a system setting		0	1.64	p = 0.001	14
Change the value of a system setting		3.86	3.21	--	14

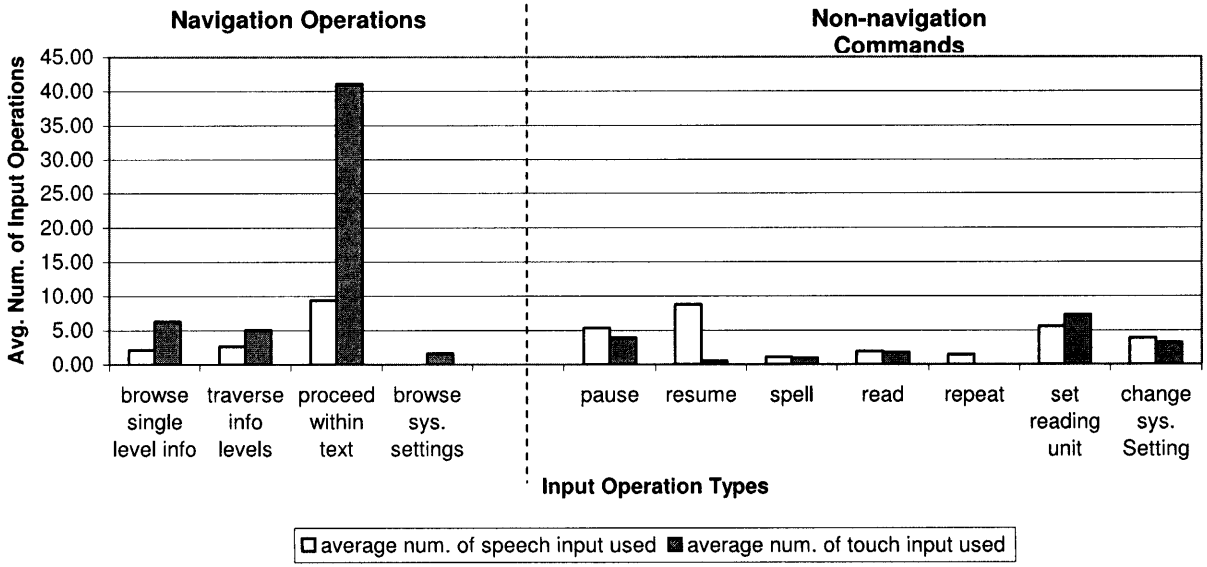


Figure 5. 3 The Operation Type-Input Mode Dependency Illustrated by Subjects' Actual Use of Input Modes

Table 5. 4 Input Mode Choice by Major Operation Categories

Operation Category	Operation Type Details	Stats of the Num. of Input Operations		Paired T Test Results
Navigation operations	Browse a single level information	Speech input: Mean: 16.768 St. Dev.: 11.729	Touchpad input: Mean: 54.071 St. Dev.: 14.362	t (13) = - 4.352 t Critical = 1.7709 p (one-tailed) = 0.0004
	Go to a different information level			
	Proceed or recede within text			
	Search for a system setting			
Subjective or verb instructions	Non-proceeding or -receding info-browsing operations within the text	Speech input: Mean: 28.571 St. Dev.: 24.861	Touchpad input: Mean: 17.857 St. Dev.: 8.848	t (13) = 1.9546 t Critical = 1.7709 p (one-tailed) = 0.0362
	Set reading unit			
	Change the value of a system setting			

5.3.2 Users' Ratings on Input Modalities for Each Operation Type

Besides looking at the amount of speech and touchpad input actually used, the experimenter also collected the subjects' opinions toward the ease of use of speech input and touchpad input, and how much the subjects liked the speech and the touchpad input (Likability), for each individual operation type. These opinions were collected during the interview session. Four questions were asked for each operation type: "how easy to use do you think speech input is when finishing this type of operation", "how much do you like to use speech input to finish this type of operation", "how easy to use do you think touchpad input is when finishing this type of operation", and "how much do you like to use touchpad input to finish this type of operation". The subjects' answers were marked on seven-point semantic differential scales, with 1 as "very easy" or "like very much", and 7 as "very difficult" or "dislike very much".

The subject's perceived ease of use of an input mode to finish a type of operation was found highly correlated to the subject's likability toward that input mode to finish that type of operation. The correlation between the perceived ease of use of speech input for an operation type and the likability to use speech input to perform that operation type was 0.9387. The same correlation between the perceived ease of use of touchpad input and the likability to use touchpad input was 0.9051.

The subjects' ratings on the ease of use of speech input and touchpad input, and their ratings on how much they like speech input and touchpad input, are found to be operation type-dependent, as well – to perform a particular type of operation, the subjects felt that a particular input mode was easier to use than the other, and that they liked one input mode better than the other for performing the operation.

The perceived ease of use and likability are elaborated along operation types in Table 5.5.

The subjects' subjective ratings reflected the patterns found in their choices of input modalities. For navigation operations, the subjects' ratings on the ease of use of touchpad input were significantly higher than those of speech input; the subjects' ratings on likability of touchpad input were significantly higher than those of speech input. For subject or verb instructions, the subjects' opinions were the opposite: the ratings on the ease of use of speech input were significantly higher than those of touchpad input; the ratings on likability of speech input were higher than touchpad input, but barely significant ($p = 0.09$, one-tailed). The detailed results of the paired t-tests are shown in Table 5.6.

Table 5. 5 Comparison of Subjective Ratings on Speech Input and on Touchpad Input along Operation Types

Operation Type		Avg. Rating of Ease of Use on Speech	Avg. Rating of Ease of Use on Touchpad	Significance (one-tailed)	Avg. Rating of Likability on Speech	Avg. Rating of Likability on Touchpad	Significance (one-tailed)	Observations
Browse a single level information		3.702	1.417	p =.00004	3.940	1.440	p =.00002	14
Go to a different information level		2.155	1.714	--	2.190	2	--	14
Proceed or recede within text		3.815	2.289	p =.00803	4.196	2.400	p =.00553	14
Non-proceeding or -receding info-browsing operations within the text	Overall	1.941	2.561	p =.050	2.131	2.653	--	14
	Pause	1.848	1.990	--	1.981	2.183	--	14
	Resume	1.470	2.835	p =.0079	1.663	3.186	p =.00675	14
	Spell	2	2.714	--	2.214	2.571	--	14
	Read the current article from the beginning	2.583	2.5	--	2.833	2.417	--	12
	Repeat	1	3.5	--	1	3.5	N/A	2
Set text unit		1.601	2.331	p =.06249	1.845	2.525	--	14
Change the value of a system setting		1.929	2.524	--	2.143	2.738	--	14

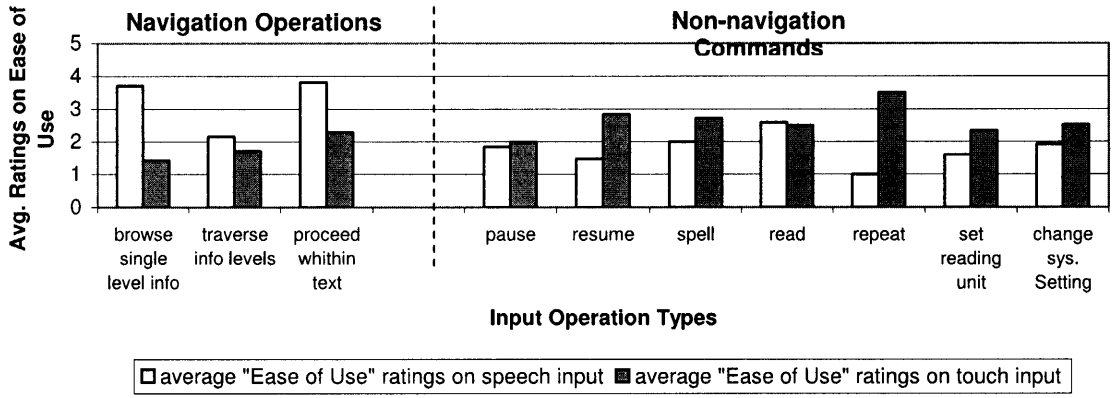


Figure 5. 4 The Operation Type-Input Mode Dependency Illustrated by Subjective Ratings on Ease of Use

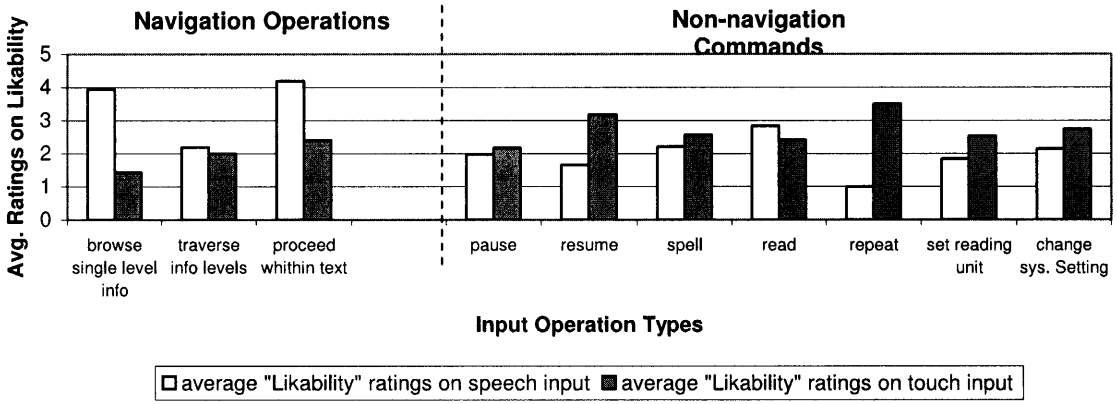


Figure 5. 5 The Operation Type-Input Mode Dependency Illustrated by Subjective Ratings on Likability

Table 5. 6 Comparison of Subjective Ratings on Speech Input and Touch Input against Operation Types

Operation Category	Operation Type Details	Subjective Ratings on Input Modes		Paired t-Test Results
		Speech Input	Touchpad Input	
Navigation operations	▪ Browse a single level information ▪ Go to a different information level ▪ Proceed or recede within text ▪ Search for a system setting	Ratings on Ease of Use:		Ratings on Ease of Use: t (13) = 3.8528 t Critical = 1.7709 p (one-tailed) = 0.001
		Mean: 3.224 St. Dev.: 1.0927	Mean: 1.806 St. Dev.:0.5785	
		Ratings on Likability:		Ratings on Likability: t (13) = 4.0454 t Critical = 1.7709 p (one-tailed) = 0.00069
		Mean: 3.442 St. Dev.:1.1695	Mean: 1.945 St. Dev.: 0.6590	
Subjective or verb instructions	▪ Non-proceeding or -receding info-browsing operations within the text ▪ Set text unit ▪ Change the value of a system setting	Ratings on Ease of Use:		Ratings on Ease of Use: t (13) = -2.0142 t Critical = 1.7709 p (one-tailed) = 0.0326
		Mean: 1.863 St. Dev.:0.7348	Mean: 2.523 St. Dev.: 1.0328	
		Ratings on Likability:		Ratings on Likability: t (13) = -1.4050 t Critical = 1.7709 p (one-tailed) = 0.0917
		Mean: 2.064 St. Dev.: 0.9343	Mean: 2.634 St. Dev.: 1.1528	

5.3.3 Input Modality Choice for Repetitive Operations

From a different perspective, input operators can be categorized as either repetitive operations or non-repetitive operations.

Two types of repetitive operations occurred during the exploratory study. One type is that when an error occurred the subject repeated the input operator until the system recovered from the error. The analysis for this type of repetition will be presented in the next section, "User Error Correction Strategies". The other type of repetition occurred when a subject needed to execute the same command repeatedly, in the same modality or in different modalities, to achieve a system action multiple times. The observation on this type of repetition is reported below.

To repeat a system action, the subjects tended to use more touchpad input than speech input. On average, each subject used 5.4 speech operations and 31 touchpad operations to perform this type of repetition. A paired t-test generates t (13) = -4.6127 and p (one-tailed) = 0.0002, which confirms that the touchpad was used significantly more than speech input. The statistics also show that most repetitions of this type took place during text comprehension and searches within text when the subjects performed reading the next or previous paragraph, sentence, or word, etc. repeatedly.

5.4 RQ3: Error Correction Strategies

The third research question is "What are users' error correction strategies on the non-visual multimodal interface?" To answer this question, the types of failed operations were identified. The subjects' error correction strategies were then elaborated.

5.4.1 Types of Input Errors

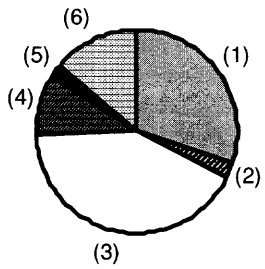
Errors took place in both speech input and touchpad input. The success rate of speech input during the experiment was 73.7%. And that of touchpad input was 95.6%. The remaining operations were unsuccessful, i.e., either partially successful (i.e., succeeded with problems, or **swp**), or completely failed (i.e., failure, or **f**).

Swp: The partially successful input operations resulted in partially successful system reactions with some execution problems. This repertoire of operations represented a situation where the system successfully interpreted and executed the input from the subject, but the information of resulted system actions was not delivered to the user clearly and fluently via the machine speech and caused confusion to the user. The unsuccessful speech feedback was mainly caused by the unnaturalness and the staccato of the machine talk generated by the Text-to-Speech engine.

F: A completely failed input operation is one of the following: (1) the system gave an explicit error message indicating that the input operation was failed, (2) the system encounted a recognition error and responded incorrectly, and (3) the system provided no respond at all to a user input.

For speech input, complete failures can be broken down to the six failure categories described in the *Failure Types* section in the "Notes Preparation and Coding" section. They are: (1) Speech output interference, (2) Speech recognition errors, (3) No-reaction symptom, (4) Environment noise interference, (5) Incorrect user mental model, and (6) Other failures, mainly caused by program bugs and execution problems. For touchpad input, the failures were (1) caused by incorrect user mental model, and (2) caused by bugs existed in the codes.

Tables 5.7 and 5.8 provide an overview of unsuccessful input operations of both speech input and touchpad input.



(1) Speech output interference	29.94%
(2) Speech recognition errors	2.55%
(3) No-reaction symptom	42.04%
(4) Environment noise interference	9.55%
(5) Incorrect user mental model	1.91%
(6) Other system failures	14.01%

Figure 5. 6 Summary of Speech Operation Failures

Table 5. 7 Summary of Success and Failures in Speech Operations

Summary Items		Count	Percentage
Total Num. of Speech Operations by All Subjects		635	100%
Successful Speech Operations		468	73.70% in total num. of speech operations
Partially Successful Speech Operations		10	1.57% in total num. of speech operations
Completely Failed Speech Operations	Total Num. of Complete Failures	157	24.72% in total num. of speech operations
	(1) Speech output interference	47	29.94% in total num. of speech failures
	(2) Speech recognition errors	4	2.55% in total num. of speech failures
	(3) No-reaction symptom	66	42.04% in total num. of speech failures
	(4) Environment noise interference	15	9.55% in total num. of speech failures
	(5) Incorrect user mental model	3	1.91% in total num. of speech failures
	(6) Other system failures	22	14.01% in total num. of speech failures

Table 5. 8 Summary of Success and Failures in Touchpad Operations

Summary Items		Count	Percentage
Total Num. of Touchpad Operations by All Subjects		1007	100%
Successful Touchpad Operations		963	95.63% in total num. of touchpad operations
Partially Successful Touchpad Operations		2	0.20% in total num. of touchpad operations
Completely Failed Touchpad Operations	Total Num. of Complete Failures	42	4.17% in total num. of touchpad operations
	(1) Incorrect user mental model	27	64.29% in total num. of touchpad failures
	(2) Other system failures	15	35.71% in total num. of touchpad failures

5.4.2 Error Correction Strategy Analysis

It was believed that when one input failed the user would switch to the other input mode to recover from the errors. However, the study disclosed different results.

Errors and correction actions were analyzed at two levels, the operator level and the error correction case level. At the operator level, the researcher observed each input operator immediately following an input failure. At the error correction case level, the researchers observed each sequence of error correction operators until the error was corrected.

Again, paired t-test was the main method for quantitative comparison. Since paired t-test is robust to non-normally distributed data, check of the normality assumption was not necessary.

5.4.2.1 Analysis at the Operator Level. In speech input, a total number of 167 unsuccessful (i.e., partially successful + complete failed) operations occurred to all fourteen subjects. Among those unsuccessful operations, 139 or 83.23% were followed by a speech input intending to overcome from the failure, and only 28 or 16.77% were followed by a touchpad input for problem fixation.

In touchpad input, a total number of 44 unsuccessful operations took place to all subjects. 37 or 84.09% of them were followed by a touchpad input and 7 or 15.92% were followed by a speech input for the purpose of failure recovery.

An unsuccessful speech input operator followed by another speech input operator to fix the problem is coded as **ss**; an unsuccessful speech input operator followed by a touchpad input operator to fix the problem is coded as **st**; an unsuccessful touchpad input

operator followed by another touchpad input operator is coded as **tt**; and an unsuccessful touchpad input operator followed by a speech input operator is coded as **ts**.

A paired t test was conducted to compare the numbers of **ss** and **st** by each subject. The paired t (13) = 4.279, with p (one-tailed) = 0.000449. It indicates that the subjects used significantly more speech input than touchpad input following an unsuccessful speech input to overcome from the preceding operation failure. Only in rare cases did the subjects switch input mode for failure recovery.

Another paired t test was conducted to compare the numbers of **tt** and **ts** by each subject. The paired t (13) = 2.760, with p (one-tailed) = 0.008114. It means that the subjects used significantly more touchpad input than speech input to fix failed touchpad operations.

The tables and charts summarizing the statistics are shown as follows.

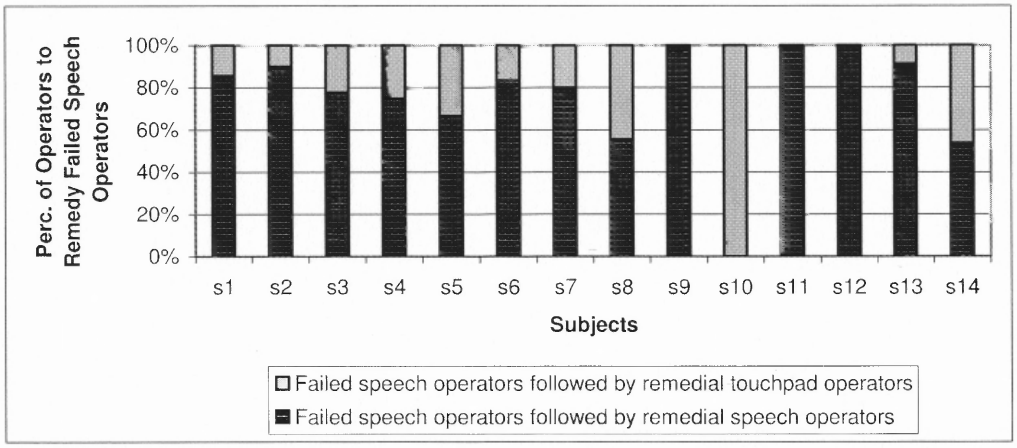


Figure 5. 7 Remedial Operators Following Failed Speech Operators

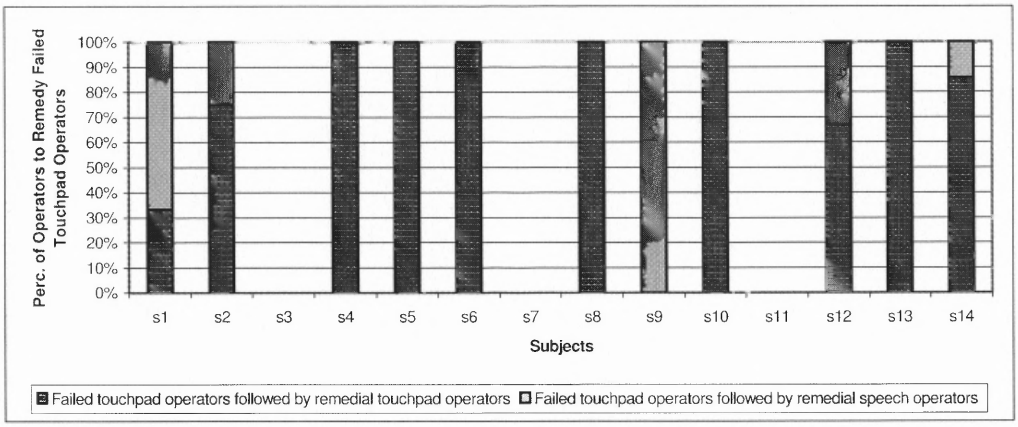


Figure 5. 8 Remedial Operators Following Failed Touchpad Operators

Table 5. 9 Remedial Operators Following Failed Speech Operators

Subject	Num. of Failed Speech Operators (FS)	Remedy Operators			
		Num. of ss*	ss% in FS	Num. of st*	st% in FS
S1	28	24	85.71%	4	14.29%
S2	20	18	90.00%	2	10.00%
S3	9	7	77.78%	2	22.22%
S4	12	9	75.00%	3	25.00%
S5	3	2	66.67%	1	33.33%
S6	12	10	83.33%	2	16.67%
S7	5	4	80.00%	1	20.00%
S8	9	5	55.56%	4	44.44%
S9	9	9	100.00%	0	0.00%
S10	1	0	0.00%	1	100.00%
S11	13	13	100.00%	0	0.00%
S12	10	10	100.00%	0	0.00%
S13	23	21	91.30%	2	8.70%
S14	13	7	53.85%	6	46.15%
Mean	**11.93**	**9.93**		**2**	
Std. Dev.	7.47	6.99	0.26	1.75	0.26
Sum	**167**	**139**		**28**	

***ss:** an unsuccessful speech input operator followed by a speech operator for failure recovery
***st:** an unsuccessful speech input operator followed by a touchpad operator for failure recovery

Table 5. 10 Remedial Operators Following Failed Touchpad Operators

Subject	Num. of Failed Touchpad Operators (FT)	Remedy Operators			
		Num. of ts*	ts% in FT	Num. of tt*	tt% in FT
S1	3	2	66.67%	1	33.33%
S2	4	1	25.00%	3	75.00%
S3	0	0	0.00%	0	0.00%
S4	10	0	0.00%	10	100.00%
S5	3	0	0.00%	3	100.00%
S6	1	0	0.00%	1	100.00%
S7	0	0	0.00%	0	0.00%
S8	3	0	0.00%	3	100.00%
S9	1	1	100.00%	0	0.00%
S10	2	0	0.00%	2	100.00%
S11	0	0	0.00%	0	0.00%
S12	6	2	33.33%	4	66.67%
S13	4	0	0.00%	4	100.00%
S14	7	1	14.29%	6	85.71%
Mean	**3.14**	**0.5**		**2.64**	
Std. Dev.	2.93	0.76	0.31	2.82	0.44
Sum	**44**	**7**		**37**	

***ts:** an unsuccessful touchpad operator followed by a speech operator for failure recovery
***tt:** an unsuccessful touchpad operator followed by a touchpad operator for failure recovery

The tables and the charts clearly show that in most cases when an operation failure occurred the subjects did not switch input mode but continued to use the same input mode for failure recovery. In the charts we can see exceptions happened to Subject 10 in speech failure recovery and Subject 9 in touchpad failure recovery. In each case only one error occurred and the subject switched input mode to remedy.

5.4.2.2 Analysis at the Error Correction Case Level. Overall, a total number of 158 errors occurred during the experiment sessions. The subjects attempted to correct 150 of them and did not try to correct the other 8 errors. Error correction operations did not always succeed at the first attempt. It took one to five attempts to correct one error. The

average attempts to correct each error were 1.32. The spread of the error correction attempts are showed in the following figure:

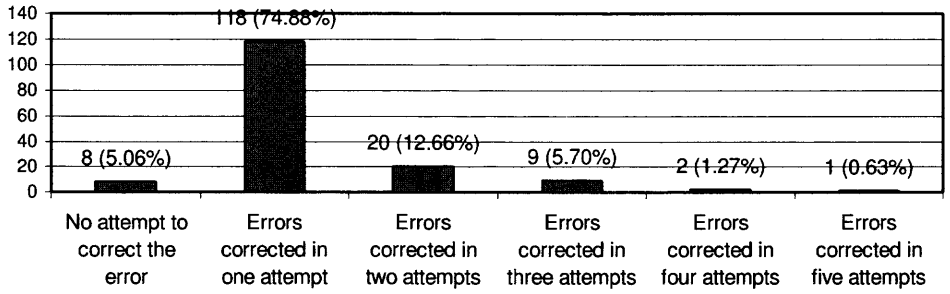


Figure 5. 9 Counts of Cases that an Error was Corrected in One, Two, Three, Four or Five Attempts

The word "case" is used for a series of operation attempts to correct one error. The error correction cases that involved input mode switches were significantly less than the cases where the error was corrected without input mode switch. A paired t (13) = -4.519 was obtained with a one-tailed p = 0.00029. This indicates that when input errors occur, users will most likely stay with the same input mode for error correction.

Table 5. 11 Counts of Error Correction Cases with & without Input Mode Switches

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Sum	%	Avg.
(1)*	6	2	2	3	1	2	1	4	1	1	0	2	2	7	**34**	22.67%	2.4
(2)*	18	10	3	5	4	9	3	8	8	2	9	8	20	9	**116**	77.33%	8.3
Sum	24	12	5	8	5	11	4	12	9	3	9	10	22	16	**150**	100%	10.7

*(1) Num. of error correction cases that involved input mode switch

*(2) Num. of error correction cases where the input mode was not switched

When an error occurs sometimes there is only one way available in the current input mode to correct the error, but sometimes multiple methods in the current input mode are available to correct the error. For example, if the speech command "pause" is not first recognized by the system, the user has to speak "pause" repeatedly until the system did recognize it. But if the speech command, "zoom in", is not recognized by the system, in speech input the user can either continue to speak "zoom in" or speak "select" to overcome the problem.

In the first situation, when only one method is available in the current input mode to correct the problem, 19 or 30.65% cases the subjects switched input mode, while in 43 or 69.35% cases the subjects did not switch input mode but used the only available method repeatedly to correct the error. A paired t-test generated t (13) = -2.604 and p (one-tailed) = 0.0109, which indicates that despite there being only one error correction method available in the failed input mode, the subjects still mostly stayed in the same mode, instead of switching the mode, for error correction. A comparison figure is showed below.

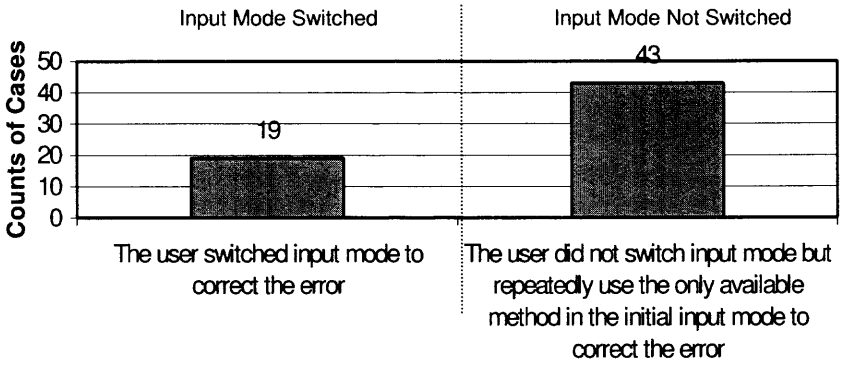


Figure 5. 10 Subjects' Error Correction Strategies When Only One Method was Available for Error Correction in the Failed Input Mode

In the second situation, when there was more than one method available for correcting the error in the current input mode, in 15 or 17.05% cases, the subjects switched input mode, while in 73 or 82.95% cases, the subjects did not switch input mode for error correction. A paired t-test confirms that input mode switch occurred significantly less than no input mode switch in this situation (t (13) = -4.833, p (one-tailed) = 0.00016). Several types of error correction actions were identified from the video analysis. They are displayed in Figure 5.11 below.

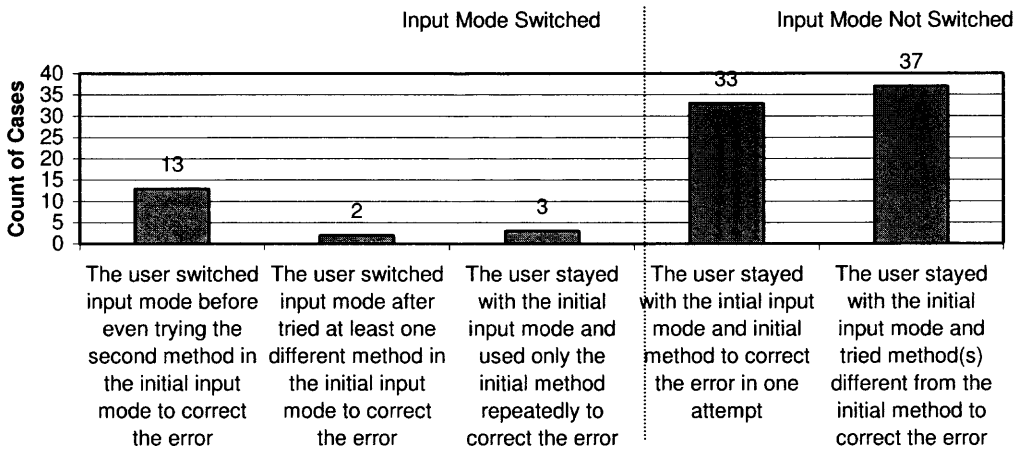


Figure 5. 11 Subjects' Error Correction Strategies When More than One Method was Available for Error Correction in the Failed Input Mode

In a total of 34 occasions input mode was switched for error correction. In 23 (67.76% of 34) cases, input mode was switched at the first error correction attempt. In six out of 34 (i.e., 17.65%) cases, input mode was switched at the second error correction attempt.

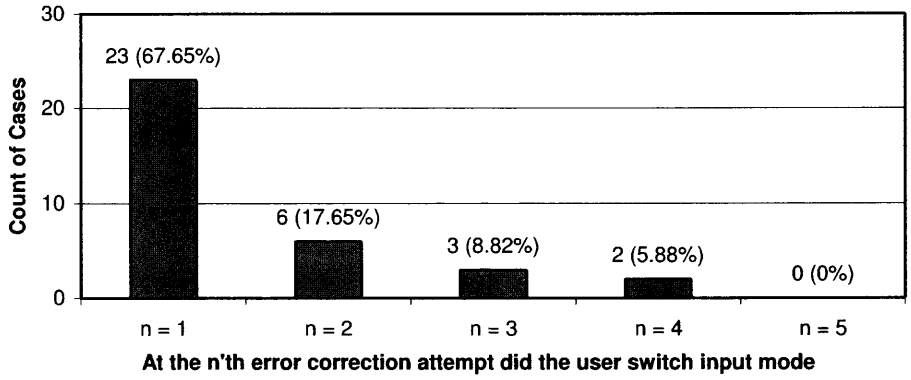


Figure 5. 12 The Time Point at Which Input Mode was Switched for Error Correction

5.5 RQ4: Effect of Training

Research question 4 is "Does training affect users' multimodal input behavior?" We asked this question because we wanted to know whether the way the training materials were presented to the subjects influenced their interaction behavior and if the effect did exist, what we should do to control the bias introduced by training during the experiment.

In the exploratory study, the training order was counterbalanced. Half of the subjects were trained for speech input first, while the other half were trained for touchpad input first. To investigate whether the training order had an effect on the subjects' choice of input method, several statistical tests were conducted.

The measures used in the tests were the amount of speech input used, the amount of touchpad input used, and the subjects' ratings on Ease of Use and Likability on each input modality. Likability refers to how much a subject liked an input modality.

Normality was checked within each corresponding data set. No test result was significant at .05, which indicated that data was normally distributed in all data sets. Therefore, parametric methods were used for all tests.

Table 5. 12 Normality Test for Modality Usage by Subjects Receiving Training in Different Orders

	Shapiro-Wilks Normality Test		
	Statistic	df	Sig.
Modality usage by people who received speech training first	0.930585	7	0.555888
Modality usage by people who received touch training first	0.979724	7	0.958187

Table 5. 13 Normality Test for Modality Ratings by Subjects Receiving Training in Different Orders

	Shapiro-Wilks Normality Test		
	Statistic	df	Sig.
Ease of use ratings on speech by subjects who received speech training first	0.873227	7	0.19804
Ease of use ratings on touch by subjects who received speech training first	0.884986	7	0.249517
Likability ratings on speech by subjects who received speech training first	0.94349	7	0.670332
Likability ratings on touch by subjects who received speech training first	0.928565	7	0.53876
Ease of use ratings on speech by subjects who received touch training first	0.945855	7	0.691884
Ease of use ratings on touch by subjects who received touch training first	0.938631	7	0.62642
Likability ratings on speech by subjects who received touch training first	0.895233	7	0.303084
Likability ratings on touch by subjects who received touch training first	0.915229	7	0.43326

First, a two-sample t test was conducted to compare the amount of speech input used by people who received speech input training first and the amount of speech input used by people who received touchpad input training first. The amount of speech input used by each subject was standardized using the formula: (num. of speech input performed by the subject) / (total num. of input performed by the subject). Thus, this percentage reflects not only the amount of speech, but also the amount of touchpad input

used by the subject. The comparison of means did not disclose any significant difference of speech and touchpad use between the two groups of people who received input trainings in reversed orders (see Table 5.14 below).

Table 5. 14 Comparing the Percentages of Speech Used by People Who Had Speech Input Training First and by People Who Had Touchpad Input Training First

Percentage of Speech Used by Subjects Who Had Speech Training First		**Percentage of Speech Used by Subjects Who Had Touchpad Training First**		**t-Test: Two-Sample Assuming Unequal Variances**		
Subject ID	Percentage	Subject ID	Percentage		*Group 1*	*Group 2*
1	66.40%	8	23.93%	Mean	0.47929	0.31343
2	35.82%	9	12.40%	Variance	0.04329	0.03501
3	11.57%	10	32.54%	Pooled Variance	0.03915	
4	55.66%	11	6.59%	Hypothesized Mean Difference	0	
5	37.84%	12	60.00%	df	12	
6	71.28%	13	45.54%	t Stat	1.56818	
7	56.93%	14	38.40%	P(T<=t) one-tail	0.07141	
Mean	**47.93%**	**Mean**	**31.34%**	t Critical one-tail	1.78229	

Second, for people who received the trainings in the same order, a paired comparison was conducted to test the difference of the number of speech input used and the number of touchpad input used by this group of people. The comparison was done twice, one for people who had speech training first, and the other for people who had touchpad training first. The results in the following Tables 5.15 and 5.16 showed that people who had touchpad training first used significantly more touchpad input during the experiment sessions. But the similar significance was not seen among the people who received speech training first.

Table 5. 15 Comparing the Amount of Speech Input and the Amount of Touchpad Input Used by People Who Had Speech Input Training First

Subject ID	Amount of Speech Input Used	Amount of Touchpad Input Used	t-Test: Paired Two-Sample for Means		
1	81	42		*Speech Used*	*Touchpad Used*
2	71	129	Mean	56.14	65.29
3	13	107	Variance	680.81	1425.6
4	58	47	Hypothesized Mean Difference	0	
5	27	46	df	6	
6	66	27	t Stat	-0.48	
7	77	59	P(T<=t) one-tail	0.33	
Mean	**56.14**	**65.29**	t Critical one-tail	1.94	

Table 5. 16 Comparing the Amount of Speech Input and the Amount of Touchpad Input Used by People Who Had Touchpad Input Training First

Subject ID	Amount of Speech Input Used	Amount of Touchpad Input Used	t-Test: Paired Two-Sample for Means		
8	27	89		*Speech Used*	*Touchpad Used*
9	15	113	Mean	34.57	78.57
10	39	85	Variance	387.29	529.29
11	5	85	Hypothesized Mean Difference	0	
12	59	40	df	6	
13	50	61	t Stat	-2.88	
14	47	77	P(T<=t) one-tail	0.01	
Mean	**34.57**	**78.57**	t Critical one-tail	1.94	

Third, the average ratings on the Ease of Use and the Likability were compared between the group that received speech training first and the group that received touchpad training first. To calculate the average ease of use rating on speech input by a group of subjects, the following formulas were used:

$$\text{Avg. Ease of Use rating on speech input by Group 1} = \frac{\Sigma \text{ Avg. Ease of Use rating on speech input by each subject in Group 1}}{\text{Total number of subjects in Group 1}}$$

$$\text{Avg. Ease of Use rating on speech input by a subject} = \frac{\Sigma \text{ Ease of Use rating on speech input to each task type by the subject}}{\text{Total number of task types}}$$

The average ease of use ratings on touchpad input, and the average likability ratings on speech input and touchpad input were calculated in similar way.

Among the comparisons of these averaged ratings, no significant difference has been found. The results are showed in the following Table 5.17.

Table 5. 17 Rating by Groups that Received Trainings in Different Orders

Items for Comparison	Ratings[1] by Group 1 [2]		Ratings[1] by Group 2 [2]		Two Sample t-Test Assuming Unequal Variance	
	Avg.	Variance	Avg.	Variance	df	p (one-tailed)
Ease of Use ratings on speech input	2.296	0.234	2.323	0.770	9	0.472
Likability ratings on speech input	2.560	0.517	2.474	0.935	11	0.426
Ease of Use ratings on touchpad input	2.575	0.864	2.037	0.446	11	0.120
Likability ratings on touchpad input	2.770	1.184	2.099	0.612	11	0.106

[1] Ratings on Ease of Use were on a seven-point scale, where 1 = very easy to use, and 7 = very difficult to use. Rating on Likability were on a seven-point scale, where 1 = like to use very much, and 7 = dislike to use very much

[2] Group 1 received speech training first, and group 2 received touchpad training first.

Fourth, for each group of people who received training in the same order, a paired comparison was conducted to look at the difference between the people's ratings on the speech input and their ratings on the touchpad input. Again, no significant difference was found.

Table 5. 18 Comparisons between Ratings on Speech Input and Ratings on Touchpad Input by Each Group that Received Trainings in the Same Order

Groups	Rating Items	Ratings on Speech		Ratings on Touchpad		Paired t-Test	
		Avg.	Variance	Avg.	Variance	df	p (one-tailed)
The group that received speech training first	Ease of Use	2.296	0.234	2.575	0.864	6	0.278
	Likability	2.560	0.517	2.770	1.184	6	0.363
The group that received touchpad training first	Ease of Use	2.323	0.770	2.037	0.446	6	0.210
	Likability	2.474	0.935	2.099	0.612	6	0.178

The following conclusions can be made on the training order effect.

(1) People who received speech training first tended to use more speech operations than people who received touchpad training first. But the tendency was not significant.

(2) People who received touchpad training first tended to use significantly more touchpad input than speech input. People who received speech training first also tended to use more touchpad input than speech input, but this tendency was not significant.

(3) People who received touchpad training first slightly tended to rate touchpad operations more positively than people who received speech training first, but the tendency was not significant.

People who received speech training first tended to rate speech operations more positively than touchpad operations. And people who received touchpad training first tended to rate touchpad operations more positively than speech operations. But none of these tendencies was significant.

5.6 Subject's Responses to the Post-Questionnaire

The subjects' overall ratings toward the AudioBrowser system were collected using the post questionnaires. Because the subjects had used (1) speech input alone, (2) touchpad input alone, and (3) mixed speech and touchpad input respectively during the three-day study, the subjects were asked to compare the three input styles.

The following table shows the statistics of the subjects' answers. Questions 1 to 6 were answered on a seven-point semantic differential scale with 1 "very easy" and 7 "very difficult". Questions 7 to 14 were answered on a five-point Likert scale with 1=strongly agree, 2=agree, 3=neutral, 4=disagree, 5=strongly disagree.

Table 5. 19 Ratings in the Post-Experiment Questionnaire

Question	Avg. Rating	Std. Deviation	Observations
Q1: Generally, how easy do you feel it is to **learn** using AudioBrowser	2.36	1.01	14
Q2: Generally, how easy do you feel it is to **use** AudioBrowser	2.5	0.76	14
Q3: How easy do you think it is to **learn** the **speech** input	2.64	1.22	14
Q4: How easy do you think it is to **use** the **speech** input	3.43	1.60	14
Q5: How easy do you think it is to **learn** the **touchpad** input	2.36	1.15	14
Q6: How easy do you think it is to **use** the **touchpad** input	2.07	1.21	14
Q7: Generally, the **mixed touchpad and speech** input is easier to use than the **touchpad** input solely	2	1.24	14
Q8: Generally, the **mixed touchpad and speech** input is easier to use than the **speech** input solely	1.57	1.02	14
Q9: I get **tired** easily when I use the **speech** input	2.79	0.89	14
Q10: I get **tired** easily when I use the **touchpad** input	3.57	1.02	14
Q11: I get **tired** easily when I use the **mixed speech and touchpad** input	3.5	1.16	14
Q12: I get **lost** easily when I use the **speech** input	2.93	1.21	14
Q13: I get **lost** easily when I use the **touchpad** input	3.36	1.08	14
Q14: I get **lost** easily when I use the **mixed speech and touchpad** input	3.79	1.31	14
Questions 1 to 6 were based on a 7-point scale with 1 "very easy" and 7 "very difficult" Questions 7 to 14 were based on a 5-point Likert scale, with 1=strongly agree, 2=agree, 3=neutral, 4=disagree, 5=strongly disagree			

To compare the subjects' ratings on speech, touchpad, and mixed input, the answers to questions 9 - 14 were standardized. The standardized values were obtained by subtracting the average rating from 5, the extreme value on the rating scale, for each question. For example, for Question 9, "I get tired easily when I use the speech input", the average rating 2.79 was subtracted from 5. The obtained value 2.21 represents the

rating on the reversed question "It is not easy to get tired when I use speech input", given 1 the most agreed extreme, and 5 the most disagreed extreme.

The ratings to questions 3 - 6 and the standardized values of the ratings on questions 9 - 14 were used for the purpose of comparison. These questions asked the subjects' opinions about ease of learning, ease of use, ease of fatigue causing (what is this?), and ease of losing orientation in the information space, of speech input, touchpad input, and mixed speech and touchpad input respectively. The lower the ratings to questions 3 - 6, the more positive the subjects' opinions were. Similarly, the lower the standardized values of the ratings to questions 9 - 14, the more positive the subjects' opinions were. The comparison results are showed in Figure 5.13 below.

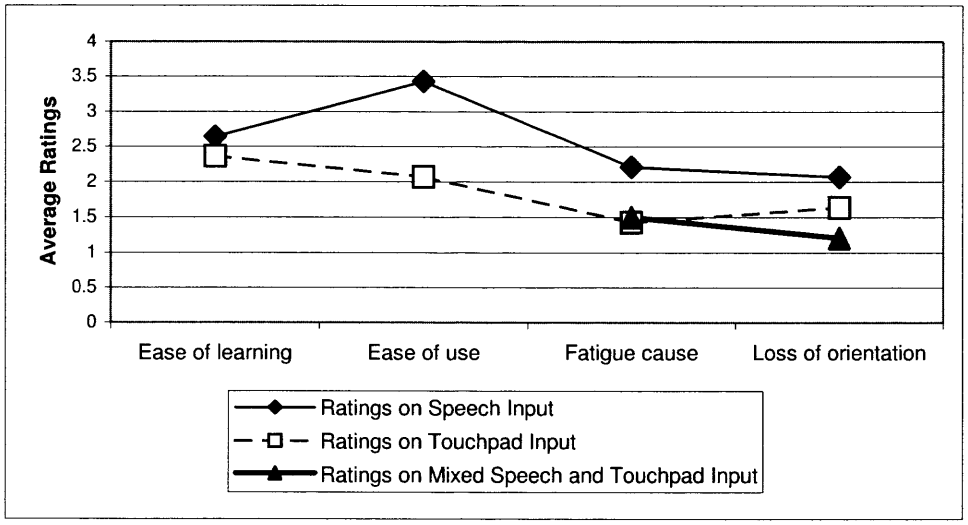


Figure 5. 13 Comparison of Subjective Ratings on Speech Input, Touchpad Input, and Mixed Speech and Touchpad Input In the Post Questionnaire

To examine possible differences within the subjects' ratings, paired t tests and ANOVA were conducted.

A paired t test comparing the subjects' answers to Q3 and Q5 showed that the ease of learning of the two input modes had no significant difference (paired t = 0.744,

one-tailed p = 0.235). A paired t test comparing the subjects' answers to Q4 and Q6 showed that the subjects felt touchpad input was significantly easier to use than speech input (paired t = 2.200, one-tailed p = 0.023).

An ANOVA test comparing answers to Q9, 10 and 11 showed that there was a tendency toward a significant difference (F = 2.497, p = 0.095). To do paired comparisons, three t tests were conducted. The paired t test comparing Q9 and Q10 showed a significant difference (paired t = -1.863, one tailed p = 0.043), indicating that the fatigue level introduced by speech input was higher than that introduced by touchpad input. The paired t test comparing Q9 and Q 11 showed similar significant difference (paired t = -1.546, one tailed p = 0.073), indicating that the multimodal input introduced a lower fatigue level than the speech input alone. The fatigue level of touchpad input and the multimodal input was not significantly different (paired t = 0.234, one tailed p = 0.409).

Another ANOVA test was conducted to compare answers to Q12, 13 and 14. The results were F = 1.775, and p = 0.183. No significance was indicated. The breakdown via paired t test resulted in a tendency of significant difference between answers to Q12 and Q14 (paired t = 1.771 and one-tailed p = 0.061), indicating a possible difference between the sense of orientation provided by the speech input alone and the combination of the speech and the touchpad inputs. No significance was found in other pairs (for Q12 and Q13, t = -0.877 and one-tailed p = 0.198; for Q13 and Q14, t = -0.921 and one-tailed p = 0.187).

Based on the subjects' experience with three input styles respectively (i.e., speech alone, touchpad alone, and mixed speech and touchpad input), it is found that the subjects

felt touchpad input was easier to learn and use than speech input, and that among the three input styles, mixed input was the least easy to cause fatigue and loss of orientation, while speech was the most easy to cause those problems.

Question 15 was a multiple-choice question with an open field, "Overall, if I have to choose one input method, I would choose: A. Speech input, B. Touchpad Input, or C. Mixed speech and touchpad input. The reason I choose it is because _____."

Eight out of 14 subjects selected mixed speech and touchpad input, 4 subjects selected touchpad input only, and 2 subjects selected speech input only. Various reasons for such choices were provided by the subjects. People who selected speech input felt that speech input was more straightforward and fun, and provided direct access to most system functions. People who selected touchpad input felt that touchpad was quicker, less taxing on memory, and less error-prone. Reasons such as some commands were more accurately recognized via touchpad than via speech input, and it required less energy to move fingers than to talk were also mentioned. The subjects who selected mixed speech and touchpad input indicated that both input modes had advantages and disadvantages; each mode was more suitable to some but not all situations; "combining the two will ensure to give the best of it". Strong reasons also include that "if I forget one command on the speech or touchpad, I probably remember it on the other", and that combined input modes provided the flexibility and accessibility that the subjects desired.

5.7 Discussion of Exploratory Study Results

5.7.1 Multimodal Use of Input Modalities

The subjects constructed mixed-mode inputs to perform the experiment tasks. Through video analysis, four reasons were identified which could have led to a total number of 222 input modes switches during the experiment. The reasons are *Change of Operation Type*, *Operation Repetition, Preceding Input Failure*, and *Start of New Task.*

Change of Operation Types: is apparently the major cause of input mode mixes. Nearly 60% of input switches were related to it. The Pearson's correlation between the number of operation type changes and the number of input mode switches reached 0.587 (p one-tailed = .005), implying that when the changes of operation types increase, the input mode switches will be more prevalent.

A typical scenario of a multimodal input due to changes in the operation type is described as follows. A user browses an article using the touchpad input. He wants to reduce the reading speed to listen to a section that is hard to understand. In reducing the reading speed, he changes the input operation type from navigation to non-navigation. In order to minimize the intervention to his navigation operation on the touchpad, he keeps his finger on the touchpad and uses the speech input to reduce the reading speed. Upon completion, he continues with the navigation operation on the touchpad.

The interpretation to this scenario is that when multitasking, users keep track of different tasks using separate input modes. In the provided instance, the subjects used the touchpad input to keep their position in the information space, so that after the intervention of the speech utterance they could efficiently continue from where the reading was interrupted.

The scenario also indicates that physical references in an input method can be actively used as additional system feedback. In the instance provided, the subjects used the touchpad to keep their temporary state in the ongoing interrupted task because the physical references provided continuous feedback that facilitated the regaining of the state after the interruption. The physical references to some extent complemented the temporal character of the auditory feedback.

Operation Repetition: also caused some input mode switches (i.e., 9.46% of the total number of switches). It has been reported before in this thesis that the subjects tended to use touchpad for repetitive operations. So when there was a need to repeat a system action, e.g., to continuously go to previous words until the word looked for was reached, the subjects tended to switch to touchpad input in case they used speech input initially.

The interpretation for the reason of this switch is that although speech is used naturally in human communication, it takes a longer time for the user to give a command than the touchpad. It is also revealed in the experiment videos that when a series of repetitive commands were given the intervals between two inputs became shorter and shorter. Speech recognition did not support faster processing of repeated commands because: (1) Repetitions did not result in faster system response time apparently to each repeated speech command. (2) When commands were repeated, the subjects' accelerated inputs were interrupted by unfinished system outputs. The subjects learned these lessons through using the system and opted to the touchpad input for repetitive operations.

Preceding Input Failure: constituted a portion of input mode switches (i.e., 13.51% of the total switches). It proved our assumption that when one input mode failed

users tend to switch to another input mode to overcome the error. The details of input mode switch due to input failures are discussed in subsection 3 in this section.

Start of New Task: or *experiment task intervention*, is the interruption in user performance due to the start of a new experiment task. When one experiment task was finished, the subject stopped to read the next experiment task. This interruption sometimes discontinued the use of one input mode, i.e., the selection of input mode for the next operation due to the effect of recency. The recency effect is defined as that the most recently used input mode tends to be used again for the next input action, because switching input mode will introduce an amount of cognitive work and is avoided by the user. 5.41% (i.e., 12 among a total of 222) input mode switches occurred with the presence of this interruption of the recency effect. For all the fourteen subjects there were a total number of 84 experiment task interventions, and 14.29% of them (i.e., 12) were accompanied with input mode switches. It shows that task reading has potentially interrupted the recency effect in the choices of input modes.

The above are four reasons that the experimenter observed that caused the subjects' input modality switches during the exploratory study. However, whether switching input modalities or not might also have been determined by the cognitive process required for switching and the cognitive resources available. It's been noticed that when doing routine tasks, such as following the instructions like "pause reading and increase the reading volume" and "spell the name of the author", the subjects switched input modalities more often and quickly. While when doing problem solving tasks, such as "summarize the reasons why the political leader's policies have failed", the subjects were more likely to execute all operators using a single input modality despite that there

were operation type changes, need of repetitive operations, and input operation failures. Further investigation is needed to reveal the effect of cognitive task types, i.e., routine cognitive tasks versus problem solving tasks, on users' input modality switch behavior.

5.7.1.1 Implications for the Design of the Controlled Experiment. The results above indicate the following:

a. The most significant impact to the subjects' choice between available input modalities was from the type of input operation. When the type of input operation changed, the subjects switched input modality to cope with the change.

b. Errors and failures in input operation influenced the subjects' modality switching. The effects on the subjects' modality switching from operation repetition and the start of new experiment tasks were relatively minor.

Therefore, in the following controlled experiment, the type of input operation should be included as an independent variable in the hypothesis testing users' input modality choices. The level of error rates, but not operation repetition or start of new experiment tasks, should be an independent variable in the hypothesis evaluating users' modality switches.

In addition, it was observed that the cognitive task types seemingly influenced the subjects' modality switching behavior, i.e., in the occasions of routine cognitive tasks, the subjects switched more frequently than in the occasions of problem solving tasks. This observation was not analyzed because the exploratory study was not designed to administer different cognitive tasks. The participants performed tasks at different cognitive levels randomly. Nevertheless, the effect of cognitive task type on users' modality switch behavior should be investigated during the experiment.

Consequently, the design of the controlled experiment will adopt the following:

(1) Types of input operation will be one of the independent variables in the hypothesis testing users' input modality choices.

(2) Input error rates will be one of the independent variables in the hypothesis testing users' input modality switching behavior.

(3) Cognitive task types will be one of the independent variables in the hypothesis testing users' input modality switching behavior.

(4) Cognitive task types should be administered as an experiment condition in the controlled experiment.

5.7.2 Input Modality – Operation Type Dependence

It has been defined that an input operator is the smallest unit of user input. A user task consists of a series of operators. The pilot study results lead to a hypothesis that there is a relationship between the input mode selected by a user and the type of the input operator under taken.

Operators belonging to the family of *navigation operations* were mostly performed using touchpad, while operators belonging to *abstract commands, or non-navigation instructions* were mostly performed using speech input.

Navigation operations involve navigating and locating information in the information space (e.g., locating a title of an article on the information hierarchy, locating a paragraph of interest in an article, and locating a system setting to adjust in the settings list). On the touchpad the user's operation could be moving a finger around on the touchpad tracks on which groups of information items and commands were arranged. The user's speech input could be saying "next article", "next paragraph", or "next setting" until the desired item is reached.

The subjects chose touchpad input to perform most navigation operations (79%). The reasons could be the following: (1) Mapping the information hierarchy onto the tracks of the touchpad allowed the transformation of a virtual information space into a two-dimensional physical space. Each item in the virtual information space, although generated dynamically, has a physical location on the touchpad and hence becomes tangible. The tangible information space assisted the subjects' formation of a mental model of the information organization. (2) Once the information structure on the touchpad was understood, the subjects used this knowledge effectively to access information quickly. The subjects did this by placing a finger onto the approximate location on a touchpad track where the desired item was likely to be. It was found through the researcher's observation that, during navigation, the subjects estimated the location of an item on the touchpad and targeted the location directly to skip unwanted information. (3) The subjects took advantage of their vision – all subjects looked at the touchpad to approximate the location they wanted to land their finger. The subjects looked at the touchpad through their task performance both when they went to explore a new route on the touchpad and when they switched back to the touchpad from using speech input.

On the contrary, speech input was not preferred by the subjects for navigation. The reasons could be the following: (1) Speech input does not provide a tangible medium to concretize the information space. Hence to comprehend the information organization through speech interaction users need to devote higher cognitive efforts than through touchpad operations. (2) A speech command (e.g., "next article") consumes longer time than a touchpad command (e.g., moving the index finger slightly on the information

browsing track) for information browsing. (3) Since speech commands were from predefined fixed grammar, they did not support naming an information item that was dynamically loaded into the system (e.g., "Go to the news aticle about WiMax phone service"). And thus, browsing using speech could only be sequential (e.g., giving the speech command "next article" repeatedly until the desired article is reached). However, sequential browsing was not preferred when the location of an item could be approximated.

Non-navigation operations or abstract commands are system commands not directly related to any spatial attributes of an information item or command item. Examples of abstract commands include pause, resume, read article, repeat, spell, set reading unit, change audio settings, etc. To issue an abstract command using the touchpad users click a touchpad button(s) or combine command searching on the touchpad tracks with button clicks. To issue an abstract command using speech users utter a speech command in the command vocabulary.

The exploratory study results show that when inputting abstract commands, the subjects tended to use significantly more speech input than touchpad input – 61% of the abstract commands were given using speech. The reasons could be twofold: (1) Uttering a speech command was more direct and hence faster than searching the command in the physical space and executing it. (2) In most cases an abstract command was given in the middle of a series of navigation commands. An instance is that when reading an article using the "next paragraph" command the subjects paused their reading and reduced the reading speed and then resumed reading. In this scenario the subjects kept track of the situations of two tasks. One was the article reading progress and the other was speed

reduction. The subjects had used the touchpad to track the article reading progress. In order not to interrupt his/her tracking, the subjects opted to speech to perform the second task.

The investigation in the subjects' choices of input modes for each subcategory of the two main input operation categories confirmed the above interpretation of the existence of an input modality – operation type dependence.

5.7.2.1 Implications for the Design of the Controlled Experiment. The following implications for the design of the controlled experiment are made.

(1) The findings continued to indicate that the operation types, i.e., navigation vs. non-navigation operations, were the major factors determining users' input modality choices. The controlled experiment, again, should include input operation types as an independent variable in the hypothesis evaluating modality choices.

(2) The results also indicated that vision was a significant advantage that the sighted subjects took to estimate the location of their wanted information on the touchpad in order to skip unwanted information quickly, and to find their way on the touchpad after modality switching. Will visually impaired users, who do not have vision to assist their use of the touchpad, make their input modality choices the same way as sighted users?

(3) The hypotheses in the controlled experiment therefore should include the level of visual impairment, i.e., low vision vs. blind, as an independent variable to evaluate the influence of residual vision on users' modality choices and modality switching behavior.

5.7.3 User Error Correction Strategies

5.7.3.1 Interpretations on Causes of Input Failures. The analysis of failures in each input mode reveals that the "no response problem" was prevalent in speech input. The symptom of the problem was no response from the system to speech input given by users. The causes of this symptom could be complicated and no sure explanation can be provided at this time. We suspect the major reason was that some background tasks

running on the operating system competed for computing resources with the speech recognition application and caused a severe delay in processing the speech recognition task. A secondary cause of the no-response problem could be the subjects' imprecise timing of the push and talk cooperation. More specifically, pushing the "push-to-talk" button about a half second before speaking a command had a higher chance of achieving a successful system response than pushing the button and speaking the command at the same time. Sometimes the subjects did the two actions simultaneously or started the utterance slightly before the button was pressed. The "no response" symptom was often observed as the result.

The incompatibility between the subjects' mental model and the system's working model was another cause of input failures in the speech and touchpad modalities. This repertoire of failures mainly related to the use of operation modes (different from the concept of input modes) in the system design. The operation modes in the system design are described as follows.

On the touchpad, the two buttons provide contextual functions as a solution to deploying all functions on the small physical design space. Each track of the touchpad is one mode. In different modes, i.e., when different touchpad tracks were touched, button clicks result in different system actions. When the top track is touched, button clicks result in zooming into or out of an information section; when the middle track is touched, button clicks result in going to the next or previous word/sentence/paragraph within a text; when the bottom track is touched, the same button clicks increase or decrease the value of an audio setting. While the subject was doing intensive information

comprehension, he/she did not always pay sufficient attention to the current system mode, which sometimes led to undesired system reactions following a button click.

For speech input, the design challenge was to find a tradeoff between two design requirements, i.e., (1) to parallel speech commands with touchpad operations to allow smooth switches between the two input modes; and (2) to avoid system modes as much as possible since the speech input was not designed on a limited design space as the touchpad input was. As a result, only a small number of speech commands are mode-sensitive. For example, when the system is in the information category browsing mode, the speech command "next" leads the system to read the title of the next information category; while when the system is in the article reading mode, "next" can lead to reading the next word, sentence, or paragraph, depending on the text unit last set by the user. Because auditory feedback indicating the current system mode was always transient and sometimes inexplicit, the subject sometimes forgot to switch to the correct system mode before giving a mode-dependent command. Consequently his/her speech input led to a system action different from his/her expectation.

5.7.3.2 Interpretations on User Error Correction Strategies. The exploratory study investigated the types of input errors that could occur in a multimodal input system that integrates speech and touch input. The exploratory study revealed the subjects' error correction behavior patterns when using a multimodal system and answered the following questions.

5.7.3.2.1 On a non-visual multimodal interface for textual information browsing, how prevalent is input modality switching following an input failure? At the operator level, immediately following a failed input, input modality switches occurred

significantly less than input operations with no modality switches. At the error correction level, in a sequence of error handling actions, input modality switches still occurred significantly less than error handling without input modality switching.

So the answer to the question is that input modality switching in the non-visual multimodal system was not a prevalent error handling strategy. This result is different from the results obtained in multimodal GUI interfaces discussed at the beginning of this article. The reasons for this different result could be (1) that the amount of mental work involved in input modality switching has prevented switching in our system (note that the user has to search for a touchpad command), (2) the lack of visual aids for our subjects meant that they could readily lose their place in the information structure if they switched modalities to correct a speech error, and (3) the availability of multiple methods for correcting an input error has encouraged the use of different methods in a single modality instead of switching modalities.

Moreover, although speech recognition had higher error rates than touchpad recognition, speech errors did not lead to higher rates of input modality switching for error correction than tactile errors. In other words, higher recognition failures do not necessarily lead to input modality switching.

5.7.3.2.1 On a non-visual multimodal interface for textual information browsing, how resistant is a user to switching input modalities when the input modality is failing? It was believed that if there was more than one method available for error correction within the same input modality, users would be more likely to use alternative methods within the same modality to correct the error, instead of switching the modality. However, we were not able to conclude this result because of the small sample size. We

did show that there was a higher tendency to switch input modalities when one input mode continued to fail.

Based on the results of this study, the error handling strategy used on multimodal GUIs and non-visual multimodal interfaces are different. Input modality switching is not likely to be a major error handling strategy on non-visual multimodal interfaces. Thus, if users should take the advantage of multimodal error correction, they will need to be taught to do so, since our study shows they are not naturally switching the modalities.

What the results suggest about the sighted subjects' error correction behavior include the following. First, it is likely that some of the speech errors that they experienced are present because of their inexperience with listening to text to speech output. It is likely that these errors will not be as prevalent with visually impaired users who are accustomed to this type of output. Second, it is possible that the resistance to modality switching may also be a result of inexperience with other aspects of our information browser. We mentioned that we suspected a higher cognitive load being a reason for this resistance to switching. This load may, in part, be a result of the subjects being inexperienced in path finding in an information space through auditory feedback. Certainly, the successful mode switching reported by Suhm et al. (2001) with the graphical user interface suggests this possibility. Finally, the sighted subjects were performing a task that required them to listen to and comprehend the information being read by the text-to-speech engine. It may be that the output modality influenced the choice of input modality, that is, the subjects stayed with the speech mode because the dialogue between the human and the computer system suggested a real conversation. Thus, the natural reaction was to continue speaking.

5.7.3.3 Implications for the Design of the Controlled Experiment. The results of the exploratory study indicate that users are more likely to fix input errors using the same input modality than switching to another modality. But when the modality continues to fail, the possibility to switch is higher.

Based on these results, the controlled experiment should test the following:

(1) Whether visually impaired users prefer to use the same input modality, instead of switching the modality, for error correction?

(2) Whether the levels of error rates influence users' error correction behavior? The corresponding hypothesis should have error rates as an independent variable and the amount of modality switches for error correction as the dependent variable. At least two levels of error rates should be administered during the experiment: high error rates and a low error rates.

(3) In addition to the impact on users' modality switches related to error correction, does the level of error rates also impact modality switches in general? This requires an investigation into not only modality switches for the error correction purpose, but also modality switches behavior in general. The hypothesis will have error rates as the independent variable, and the amounts of the two types of modality switches as the two related dependent variable.

5.7.4 Training Order Effect (Primacy Effect)

We predicted that the input modality learned first would be the primary modality used later (the primacy effect). The study results showed a small primacy effect in the subjects' use of the input modalities. People who received speech training first tended to use more speech and less touchpad operations than people who received touchpad training first. This tendency was nearly significant (two sample $t(12)=1.568$, $p=0.07$). People who received touchpad training first tended to use more touchpad input than speech input. This tendency was significant (paired $t(13)=2.88$, one-tailed $p=0.01$), while people who received speech training first also tended to use more touchpad input than

speech input, but this tendency was not significant (paired t(13)=0.48, one-tailed p=0.33). However, the training order of different input modes did not significantly affect the subjects' subjective ratings of different input modalities. A design implication can be derived from these results: since there is a clear advantage in choosing an input modality that matches an operation type, user-training materials should be designed so that the more advantageous input modality is taught first for performing a type of task.

5.7.4.1 Implications for the Design of the Controlled Experiment. To limit the number of factors impacting the subjects' input modality choice, the training order effect should be controlled during the experiment. Rather than administering the training materials in speech input and touchpad input separately and counterbalancing the training orders, the experimenter should provide multimodal input training, i.e., mixing the speech and touch input training, and guiding participants to practice multimodal input during the training.

5.7.5 Other Interpretations

During the experiment sessions, the subjects performed a total number of 1642 input operators. Among them, 39.04% were speech input and 60.96% were touchpad input. The following reasons might have contributed to more use of touchpad than speech. (1) Touchpad input is more robust than speech input. The success rates of touchpad input and speech input were 95.63% and 73.70% respectively. (2) When using the touchpad the sighted subjects could see the touchpad tracks and approximate the location of an item on them. (3) Overall, the subjects performed more navigation operations than non-navigation instructions (954 (58%) vs. 688(42%)). The subjects commonly chose touchpad input for

navigation and speech for non-navigation commands. These reasons have possibly caused more use of touchpad input than speech input during the experiment.

Through the analysis of the subjects' answers in the post questionnaire, it was found that although the speech input was not more difficult than the touchpad input to learn, the speech input was more difficult than the touchpad input to use. It was also found that speech input alone caused significantly higher level of fatigue in use than either the touchpad input alone or the mixed use of speech and touchpad inputs. These results could be the consequences of the same reasons mentioned above in (1), (2) and (3). However, it is interesting to see that the touchpad input did not provide significantly better support in orientation in the information space than the speech input, while the mixed speech and touch provided significantly better support in information space orientation than the speech input. The explanation could be that the touchpad input provided certain advantages in keeping track of a user's location when the user navigates the information space, while the speech input allowed the user to deal with interrupting tasks separately without leaving the current navigation on the touchpad. Resuming the navigation from the interrupted point became easier because of these separate processes. Hence the combined speech and touch provided a better sense of orientation in the information space than the speech alone, while the touchpad input alone did not.

5.7.5.1 Implications for the Design of the Controlled Experiment. Since most participants looked at the touchpad when executing touchpad operations, a hypothesis could be constructed to predict that vision has provided additional advantage in path finding on the physical space, and hence users with any usable vision will be more likely to use the touchpad than users with no working vision. This difference between users

with low vision and users with no vision should be investigated during the controlled experiment.

5.8 Summary of Exploratory Study Results and Implications for Design of Controlled Experiment

In summary, the analysis and discussion suggest the following findings from the exploratory study. They also suggest the consequent changes or refinement in the research questions, as well as the hypotheses to test in the controlled experiment. The following sections concisely summarize these points.

5.8.1 RQ1

"When interacting with a non-visual multimodal system, do users use multimodal or unimodal input?"

5.8.1.1 Major Observations. All sighted subjects chose to use multimodal input, rather than single input modality. Reasons for switching between input modalities varied, but the most prominent reasons are the *change of operation types* and the *occurrences of input errors*.

In addition, although not analyzed due to the lack of control in the exploratory study, the cognitive task types seemingly influenced the subjects' modality switching behavior, i.e., when performing routine cognitive tasks, the subjects switched input modalities more frequently than when performing problem solving tasks.

5.8.1.2 Implications for the Controlled Experiment. The following implications were obtained.

(1) Types of input operation should be one of the independent variables in the hypothesis testing users' input modality choices.

(2) Input error rates should be one of the independent variables in the hypothesis testing users' input modality switching behavior.

(3) Cognitive task types should be one of the independent variables in the hypothesis testing users' input modality switching behavior.

(4) Cognitive task types should be administered as an experiment condition in the controlled experiment.

5.8.1.3 Hypotheses. Based on each of the implications above, the following hypotheses can be constructed:

(1) When performing navigation operations, users will use significantly more touchpad input and less speech input than when performing non-navigation operations.

(2) When error rate increases, users will switch input modality significantly more frequently for error correction.

(3) When performing routine cognitive tasks, users will switch input modality significantly more frequently than when performing problem solving tasks.

5.8.2 RQ2

The original research question was: "If users choose to use multimodal, rather than unimodal input, do they have special multimodal input patterns – i.e., is there a relationship between the type of input operation and users' choice of input modality?"

5.8.2.1 Major Observations. The observations continue to indicate that input operation types are the major factor influencing users' input modality choices. The observations also infer that visually impaired users' level of usable vision might influence

whether they prefer touchpad input or not, because all sighted subjects used their vision to help with locating their fingers and way-finding on the touchpad.

5.8.2.2 Implications for the Controlled Experiment. In combination with the indications from RQ1, RQ2 should be modified to accommodate more factors potentially influencing users' multimodal input patter. RQ2 is therefore modified as follows: "Do any of the following factors have an impact on visually impaired users' multimodal input usage: type of input operator, level of visual impairment, and type of cognitive task?"

5.8.2.3 Hypotheses. Among visually impaired users, users with working vision will use the touchpad input more frequently than users with no working vision.

5.8.3 RQ3

The original research question was: "What are users' error correction strategies on the non-visual multimodal interface?"

5.8.3.1 Major Observations. Sighted subjects mostly used the same input modality, rather than switching to the other input modality, to correct errors and failures.

5.8.3.2 Implications for the Controlled Experiment. The controlled experiment should continue to investigate whether this behavior pattern is the same as visually impaired users' error correction strategy.

In addition, the observation discussed in RQ1, that the higher error rates increased users' input modality switched, should be incorporated into RQ3. RQ3 should address whether users' error correction strategy and multimodal input pattern change with variation of error rates.

Therefore, RQ3 is revised to: "Will errors change users' multimodal interaction behavior?" It consists of the following more specific and detailed research questions:

- RQ3a. How do visually impaired users correct errors, by switching input modalities or not?
- RQ3b. Whether different levels of error rates influence users' error correction behavior?
- RQ3c. Does variation of error rates affect only users' error correction related modality switches, or users' modality switching behavior in general?"

Furthermore, the previous observation that sighted subjects all used their vision to help with approximating the location of the wanted information and way-finding on the touchpad suggested that two more detailed research questions could be added:

- RQ3d. Whether the level of visual impairment has an influence on users' modality switches for error correction?
- RQ3e. Whether the level of visual impairment has an influence on users' modality switches in general?

5.8.3.3 Hypotheses. The hypotheses should test the following:

(1) When errors occur, instead of switching to another input modality, visually impaired users will be more likely to continue to use the failing modality for error correction.

(2) The same as a hypothesis in RQ1, When error rate increases, users will switch input modality significantly more frequently for error correction.

(3) When error rate increases, users will switch input modality significantly more frequently in general.

(4) Among visually impaired users, users with working vision will switch input modalities more frequently for error correction than users with no working vision.

(5) Among visually impaired users, users with working vision will switch input modalities more frequently in general than users with no working vision.

5.8.4 RQ4

RQ4 addressed the effect or training order: "Does training affect users' multimodal input behavior?"

5.8.4.1 Major Observations. There was a small training order effect that had influenced the subjects' choice of input modalities. The input modality taught first became the primary modality the subjects used.

5.8.4.2 Implications for the Controlled Experiment. To limit the number of independent variables, the training order effect should be controlled during the experiment. The way to control it is to teach the participants to use multimodal input, rather than speech input and touchpad input separately. By administering multimodal input tutorials, the training order effect should be eliminated. RQ4 therefore should be removed from the controlled experiment.

CHAPTER 6

DESIGN OF CONTROLLED EXPERIMENT WITH VISUALLY IMPAIRED USERS

6.1 Overview

The goal of the controlled experiment was to test the hypotheses formed from the exploratory study with visually impaired users and to answer the research questions.

The experiment procedure was designed based on but changed from the procedure of the exploratory study. The major changes are listed below. These changes were also made to minimize influences of factors not within the scope of evaluation on the subjects' choice of input modalities.

- Keyboard control was implemented into the AudioBrowser system as a Wizard of Oz method to allow visually impaired participants to use the speech recognition software for speech input and to allow the experimenter to manipulate input errors. The Wizard of Oz method allowed the experimenter to generate system output and errors using the keyboard without the participants' awareness during the experiment sessions. This implementation controlled the influence of speech recognition failures on the subjects' choice of modalities.

- The user training materials were modified. The participants were trained to use multimodal input rather than the two modalities separately. This modification controlled the training order effect on the subjects' choice of modalities.

- The participants were tested on their ability to understand computer synthesized speech output before participating in experiment sessions. During the test each participant used the reading speed that he or she felt most comfortable with to finish a listening comprehension test designed by Educational Testing Services (ETS). This test established a baseline to help to understand whether the subjects' ability to comprehend synthesized speech affected their choice of input modalities.

- The length of the experiment was decreased from about seven hours over three days to about five hours over two days to reduce the participants' fatigue. This change reduced the effect of fatigue on the subjects' choice of modalities.

In addition, the recruiting procedure filtered applicants with less than one year computer use experience and applicants younger than 20 years old or older than 60 years old. These recruiting criteria limited potential influences from the subjects' background and demographics on their choice of input modalities.

The experiment tasks were designed to incorporate routine cognitive tasks and problem solving tasks. Error rates were controlled so that each subject experienced a session with human errors only and a session with increased error rates by having both human errors and artificially generated (system) errors. Human errors refer to mistakes made by users. System errors refer to the system's failure to process user commands. In this experiment, fixed errors were introduced in the second session using the Wizard of Oz method in every participant's experiment session.

The next section of this chapter describes the research questions and hypotheses derived from the exploratory study, the subjects recruited, the AudioBrowser system modified for controls and manipulations, the administration of experiment conditions, the procedures carried out, and the tasks given to the subjects.

6.2 Revised Research Questions

To confirm and deepen the understanding of the results obtained from the exploratory study with sighted users, an experiment with visually impaired users was conducted. Some modifications and refinements in the research questions were made based on the rationale from the previous chapter.

In the exploratory study, the phenomena of interest were the use of multimodal input, the influences of types of input operators on choices of input modality, and

multimodal error correction strategies. In addition, the effect of training order was examined.

In the confirmatory experiment with visually impaired users, the previous factors of interest remain, but the training order was treated as a controlled variable – the subjects were trained to use multimodal input, instead of one input modality after another. The reason for doing this was to eliminate the training order effect and improve the precision of the results.

Besides, two new factors of interests emerged and were included in the experiment. They are the level of visual impairment and the type of cognitive task.

The level of visual impairment was included because through conversation and observation, it was found that people with any vision at all attempted to use their vision whenever they could. Being able to see the touchpad might obtain additional spatial cues that could facilitate the use of the touchpad and hence encourage these subjects to use the touchpad more often. The level of visual impairment, therefore, is within the scope of investigation on users' multimodal input patterns.

The type of cognitive task was included because based on working memory theories, people are generally better at dividing attention across modalities than within a single modality; and that different cognitive tasks might result in different attention divisions that influence users' use of the multimodal interaction system.

The research questions for the confirmatory experiment are therefore as follows:

- **RQ1**: When interacting with a non-visual multimodal system, do visually impaired users use multimodal or unimodal input?

- **RQ2**: Do any of the following factors have an impact on visually impaired users' multimodal input usage: type of input operator, level of visual impairment, and type of cognitive task?

- **RQ3**: Will errors change users' multimodal interaction behavior?

  RQ3 further embraces three detailed research questions:

  - **RQ3.1**: Do users switch input modalities when correcting errors?
  - **RQ3.2**: Will level of error rates change users' error correction strategies?
  - **RQ3.3**: Will level of error rates influence users' overall modality switching patterns?

- **RQ4**: Can we conclude any common or different multimodal interaction patterns between sighted users and visually impaired users?

6.3 Hypotheses

The exploratory study results have indicated the way the hypotheses should be constructed.

Testing models and hypotheses are listed in Table 6.1.

Table 6. 1 Research Questions, Quantitative Models and Hypotheses

<table>
<tr><th>RQ</th><th colspan="2">Models and Hypotheses</th></tr>
<tr><td>RQ1</td><td colspan="2">When interacting with a non-visual multimodal system, do visually impaired users use multimodal or unimodal input?</td></tr>
<tr><td>RQ2</td><td colspan="2">Do any of the following factors have an impact on visually impaired users' multimodal input usage: type of input operator, level of visual impairment, and type of cognitive task?</td></tr>
<tr><td></td><td colspan="2">Model 2.1: Effects of level of visual impairment and type of operator on users' choice of input modality
Independent variables:
• Level of visual impairment (with working vision vs. without working vision)
• Type of input operator (navigation operator vs. non-navigation operator)
Dependent variable:
• Choice of input modality (speech input vs. touch input)</td></tr>
<tr><td></td><td>H2.1:</td><td>Visually impaired users' choice of input modality is determined by the input operator type being executed, and users' level of visual impairment.</td></tr>
<tr><td></td><td>H2.1a:</td><td>When performing navigation operations, users will use significantly more touchpad input and less speech input than when performing non-navigation operations.</td></tr>
<tr><td></td><td>H2.1b:</td><td>Visually impaired users with working vision will use the touchpad input significantly more than users with no working vision.</td></tr>
<tr><td></td><td colspan="2">Model 2.2: Effects of cognitive task types on input modality switches
Independent variable:
• Cognitive task type (Routine Cognitive Tasks vs. Problem Solving Tasks)
Dependent variable:
• Frequency of input modality switches</td></tr>
<tr><td></td><td>H2.2:</td><td>When performing routine cognitive tasks, users will switch input modality significantly more frequently than when performing problem solving tasks.</td></tr>
</table>

(Continued)

RQ	Models and Hypotheses	
RQ3	Will errors change users' multimodal interaction behavior?	
RQ3.1	**RQ3.1:** Do users switch input modalities when correcting errors?	
	Model 3.1: Whether users are more likely to switch input modalities or use the same input modality to correct errors? Independent variable: • Type failing modality (speech failure vs. touch failure) Dependent variable: • Error correction strategy (correcting by switching modality vs. correcting within modality)	
	H3.1	Users will correct errors in the failing modality significantly more often than correcting them in another modality.
RQ3.2 & 3.3	**RQ3.2:** Will level of error rates change users' error correction strategy? **RQ3.3:** Will level of error rates influence users' overall modality switching pattern?	
	Model 3.2 & 3.3: The effects of error rate and level of visual impairment on modality switches for error correction and users' overall modality switches Independent variables: • Level of visual impairment (with working vision vs. without working vision) • Error rate (low error rate vs. high error rate) Dependent variables: • Error correction strategy (correcting by switching modality vs. correcting within modality) • Total amount of modality switches	
	H3.2 & 3.3:	Users' modality switches for error correction and modality-switching behavior in general are determined by the level of error rates and users' level of visual impairment.
	H3.2 a	Users with working vision will switch input modalities more frequently for error correction than users with no working vision.
	H3.2 b	When error rate increases, users will switch input modality significantly more frequently for error correction.
	H3.3 a	Users with working vision will switch input modalities more frequently in general than users with no working vision.
	H3.3 b	When error rate increases, users will switch input modality significantly more frequently in general.
RQ4	Can we conclude any common or different multimodal interaction patterns between sighted users and visually impaired users?	

6.4 Experiment Design

The experiment was a factorial design, which incorporated three within-subject variables and one between subject variable.

- The independent variables include the following:
  - Level of visual impairment (with working vision vs. without working vision) – between-subject variable
  - Type of input operator (navigation operator vs. non-navigation operator) – within-subject variable
  - Type of failing modality (speech failure vs. touch failure) – within-subject variable
  - Cognitive task type (Routine Cognitive Tasks vs. Problem Solving Tasks) – within-subject variable
  - Error rate (low error rate vs. high error rate) – within-subject variable
- The dependent variables include the following:
  - Choice of input modality (speech input vs. touch input)
  - Frequency of input modality switches
  - Frequency that each error correction strategy was used (correcting by switching modality vs. correcting within modality)

    The variables for each testing model are stated in Table 6.2.

Table 6. 2 Experiment Design for Testing Models

Model 2.1: Effects of level of visual impairment and type of operator on users' choice of input modality Experiment design: • 2x2 factorial design with one between subject variable and one within subject variable Independent variables: • Between-subject: Level of visual impairment (with working vision vs. without working vision) • Within-subject: Type of input operator (navigation operator vs. non-navigation operator) Dependent variable: • Choice of input modality (speech input vs. touch input)
Model 2.2: Effects of cognitive task types on input modality switches Experiment design: • Single factor design Independent variable: • Within-subject: Cognitive task type (Routine Cognitive Tasks vs. Problem Solving Tasks) Dependent variable: • Frequency of input modality switches
Model 3.1: Whether users are more likely to switch input modalities or use the same input modality to correct errors? Experiment design: • Single factor design Independent variable: • Within-subject: Type of failing modality (speech failure vs. touch failure) Dependent variable: • The frequency of error correction with input modality switching and the frequency of error correction without modality switching
Model 3.2 & 3.3: The effects of error rate and level of visual impairment on modality switches for error correction and users' overall modality switches Experiment design: • 2x2 factorial design with one between subject variable and one within subject variable Independent variables: • Between-subject: Level of visual impairment (with working vision vs. without working vision) • Within-subject: Error rate (low error rate vs. high error rate) Dependent variables: • Error correction strategy (correcting by switching modality vs. correcting within modality) • Total amount of modality switches

In the experiment sessions the subjects performed news article reading tasks. They were told they could use the speech and touch input any way they wanted. For each condition, the subjects' choices of input modalities, input modality switches, and error correction strategies were collected.

The experiment used a Wizard of Oz feature built in the AudioBrowser system which allowed the experimenter to generate system output using the keyboard which the subjects were not aware of. The Wizard of Oz feature was used for two purposes: to avoid the inaccessibility problems with the speech recognition system and to control error rates.

Different from the training sessions in the first experiment that taught speech input and touch input separately, the training sessions of the second experiment acquainted the participants with speech and touch input simultaneously. This change eliminated the small training order effect found in the first experiment.

To get a baseline on the subjects' ability to understand computer-synthesized speech output, the subjects participated in a listening comprehension test. Each subject read two articles using AudioBrowser and then answered a series of reading comprehension questions on the articles listened to. In order to answer the questions correctly each subject had to understand the articles read using the synthesized speech. The articles, questions and standard answers were from the Listening Comprehension Section of TOEFL tests (Test of English as a Foreign Language) published on ETS' web site (ETS, 2005). None of the subjects had read these articles before. The range of the test scores indicated acceptable abilities of the subjects to understand computer-synthesized speech output.

The following sections describe the experiment conditions, the subjects, the experiment system, and the procedure.

6.4.1 Experiment Conditions

In this section the methods used to administer the independent variables are described in more details.

6.4.1.1 Cognitive task types (routine cognitive tasks vs. problem solving tasks). Routine cognitive tasks are tasks using routine cognitive skills. A routine cognitive skill is one where the person executing the skill has the correct knowledge of how to perform the task and simply needs to execute that knowledge (Card et al. 1980 and 1983; John and Kieras, 1996). In other words, when doing routine cognitive tasks, users are no longer problem solving, but rather applying procedural knowledge to a relatively familiar task or a routine procedure.

A training session and two practice sessions, totaling two hours across two days, were designed and administered to ensure that the participants obtained the required routine cognitive skills in using AudioBrowser prior to the experiment session.

Some examples of routine cognitive tasks used in the experiment are:

- Enter the Times Magazine Special Issue Section.
- Set the reading unit to Sentence.
- Read three sentences and pause reading.

Problem solving tasks, in contrast, are tasks requiring the modulation and control of routine cognitive skills without prepared procedural knowledge (Goldstein and Levin,

1987). When users create their own method to achieve a given goal, they are doing a problem solving task.

An example of the type of problem solving task used in the experiment is:

- You will read an article titled *The People Who Influence Our Lives* and answer questions based on the article. Find the article and the questions in the *Times Magazine Special Issue* Section. You can revisit the article as many times as you need to look for answers to the questions.

6.4.1.2 Input error rates (low error rates vs. high error rates). The Wizard of Oz feature of the AudioBrowser system was used to control the level of error rates in both the speech input and the touch input.

There were two types of errors. One type was generated when the subjects used a wrong command or touched the touchpad accidentally. These are called human errors. The other type was generated when the system processed user commands incorrectly. These are called system errors. The Wizard of Oz method was used to simulate both human errors (e.g., touching the touchpad accidentally) and system errors (e.g., recognizing a speech command incorrectly).

In the condition with low error rates, the experimenter did not administer errors. The error rates reflect the human errors the participants made. In this condition, the error rate ranged from 3.35% to 24.45%. The mean and standard deviation of the subjects' error rates were 12.68% and 6.18%.

In the condition with high error rates, in addition to human errors, the experimenter administered system errors during the subjects' speech and touch input using a keyboard control. The subject was not aware of this manipulation. The error rate,

in this condition, ranged from 14.63% to 32.47%. The mean and standard deviation were 26.35% and 4.70%.

For each subject, the difference between the error rates of the two conditions ranged from 5.19% to 25.91%, averaged at 13.67% with a standard deviation of 6.43%. In other words, the participants all encountered an increase in input errors when proceeding from the occasional failure condition to the frequent failure condition.

6.4.1.3 Types of input operators (navigation operators vs. non-navigation instructions). One goal of the experiment was to measure the input choices of users of a multimodal system in the performance of different tasks. Because of this goal, it was not possible to control the operations (navigation vs. non-navigation) users would use in their execution of the tasks. Hence the operation type is a random independent variable reflecting each user's personal problem solving path. The operation types were collected during subjects' task performance sessions, then used as a factor to predict the input modalities the users would choose.

6.4.1.4 Level of visual impairment (with working vision vs. with no working vision). The level of visual impairment was incorporated as an independent variable because people with any vision will depend on their vision as much as possible, as such, this dependence might have an impact on user's use of different input modalities because touch uses more of a person's visuo-spatial skills.

During the research, the users were interviewed about their backgrounds first, including questions about their current vision that indicated whether they had working vision or not. The users were then divided into two groups for data analysis based on their visual ability.

6.4.1.5 Administration of experiment conditions. Because the increased error rate condition is an extreme condition that could affect other factors in the experiment, these factors were not administered in this condition.. For example, cognitive task types were administered only within the condition of low error rates.

Table 6. 3 Experiment conditions

Low error rates				High error rates
Routine cognitive tasks		Problem solving tasks		
Navigation operators	Non-navigation instructions	Navigation operators	Non-navigation instructions	

6.4.2 Subjects

Subjects were recruited at an annual convention in New Jersey hosted by the National Foundation for the Blind. The convention was aimed at introducing technologies that assisted visually impaired people. At the convention the AudioBrowser research group set up a booth to demonstrate this research and software. The group distributed flyers describing the research, which were printed in both Braille and large type, to the convention attendees. The faculty researcher in the group gave a workshop in how HCI research had been making differences to assist the life of visually impaired people. During the exhibition and the workshop, visually impaired attendees were invited to try using AudioBrowser to read news articles published on that day. People who were interested were encouraged to leave their contact information for participation in the experiment. They were also encouraged to refer their visually impaired friends to participate in the research.

Subjects were screened based on the following standards:

- The subject had his / her visual impairment for at least three years.
- The subject used a computer for at least one year.
- The subject used synthesized computer speech output for at least one year.
- The subject was over twenty but less than sixty years old.

A total of twenty subjects who met the above criteria were recruited. Of them ten were female and the other ten were male. However, one male subject did not finish the experiment – he constantly fell asleep during the experiment. It was found that he was suffering from a brain tumor that caused this narcolepsy. His data was therefore eliminated from the data analysis.

6.4.3 Experiment System

During the experiment, the AudioBrowser system was running on a Dell Inspiron laptop. The laptop was placed in front of the experimenter, so that the experimenter could monitor the user's touchpad interaction through AudioBrowser's visual output on the screen and control the system using the keyboard. To maximize the quality of the auditory output, a pair of speakers were connected to the laptop and placed to face the subject. A touchpad connected to the laptop was given to the subject. A microphone for speech input was placed in front of the subject but was not connected. Instead, the experimenter, unbeknownst to the subject, typed in all spoken commands. A video camera was set up on a tripod to videotape the user's interaction. Figure 6.1 illustrates this system setup.

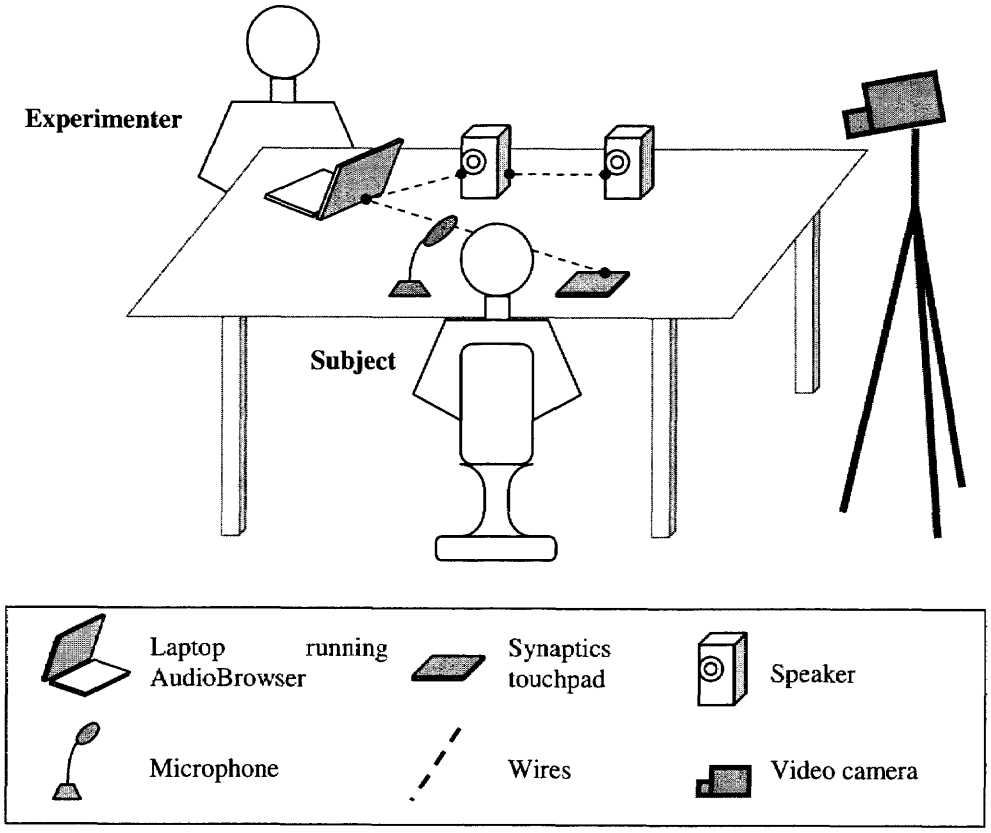


Figure 6. 1 The Experiment System Setup

6.4.3.1 Control Using Wizard of Oz Method. At the time the speech recognition input was built into the AudioBrowser system, the success of available and affordable speech recognition engines all relied on user profile-based speech training. Users had to train the speech recognition engine individually by reading texts displayed on the screen and repeated words unrecognized by the engine until all user utterances were recognized. During the training the speech engine built a user profile for each user. When a user wanted to use speech input s/he turned on her/his profile. The longer a user trained the engine the higher the recognition rate the engine returned for the user during her/his later use. This user profile-based engine training method is commonly used in low-cost or free

speech engines which are likely to be used in accessibility software. The availability of low-cost speech understanding systems is changing so that this issue is disappearing.

However, visually impaired users are excluded from any speech recognition system that requires profile-based speech engine training because screen readers are typically not adapted to these systems which display visually what is to be read and then dynamically indicate those passages that need to be reread during the training.

In order to simulate a situation where the speech recognition system works for visually impaired users, a Wizard of Oz feature was built into the AudioBrowser system. The Wizard of Oz feature allows user controls via a standard keyboard – when an utterance is given by a subject, the experimenter can generate system responses accordingly via keystrokes. All commands and system speech output including error messages were mapped onto the keyboard that was controlled by the experimenter. The experiment participants were therefore not required to press the "push-to-talk" button before giving a speech command. This Wizard of Oz feature also had the distinct advantage of controlling speech recognition errors in the experiment.

Some examples of the map between speech commands and keyboard keys are:

- Ctrl + S = the speech command "next sentence"
- Alt + S = the speech command "previous sentence"
- F2 = An error message for an invalid speech input, "Sorry, this command is invalid"

When the user gave a speech command, the experimenter hit the corresponding keys, and the system responded by repeating the speech command first, then executing the command.

The following picture shows a map of speech output and system commands on the keyboard. A total of 56 speech commands were available through single or combined keystrokes.

Figure 6. 2 Keyboard Used to Control AudioBrowser's Wizard of Oz Feature

Besides speech commands, keystrokes were also able to simulate touchpad commands. The third row of keys from *Q* to *}*, the fourth row from *A* to *"*, and the fifth row from *Z* to *?* on the keyboard correspond to the top, middle and bottom tracks of the touchpad. Each key correspond to a segment on a touchpad track. For example:

- One stroke on the key *Q* equals a touch on the first segment of the top touchpad track.
- One stroke on *W* equals a touch on the second segment of the same track.
- One stroke on *A* equals a touch on the first segment of the second touchpad track, etc.

The plus key and the minus key correspond to the right and left zoom buttons on the touchpad. In addition, a key is used to disable the touchpad.

Thus, the Wizard of Oz feature not only simulates the situation where the speech recognition system "works" for visually impaired users, but also allows generating "errors" in users' speech and touch inputs.

6.4.3.2 Articles Read by AudioBrowser. The articles that AudioBrowser read for the subjects during the experiment came from several resources. The articles used for the AudioBrowser tutorial and practice sessions were from an earlier New York Times issue. The articles used during the experiment sessions were from the New York Times, the Times Magazine, and the listening comprehension test from TOEFL tests created by ETS.

Due to undesired interpretations of the Microsoft Text to Speech (TTS) engine, all articles were proofread before being used for the experiment. Changes to the articles were made whenever necessary in order to assure the correctness of the speech by the TTS engine. For example, some acronyms and punctuations could not be interpreted correctly by the TTS engine and so were modified. Some examples of the modifications are replacing Dr. with Doctor or Drive, replacing CA, Cal., and Calif. with California, replacing U.S. with U S, etc. For words that the TTS engine could not pronounce correctly, the experimenter tweaked the spelling so that the TTS engine could provide a pronunciation closer to the word used in the article.

Tables 6.4 and 6.5 give a breakdown of the two sets of articles read by AudioBrowser during the experiment. Figure 6.3 presents the hierarchical structure of reading set 2 that was used in AudioBrowser.

Table 6. 4 Article Set One Read by AudioBrowser

Purpose	Used by the participants for the training and practice sessions
Resource	To assure that the users were not familiar with the articles, the experimenter chose old New York Times articles published in the year 2000
Size of article set	A total of eight basic news categories that consist of 25 subsections and 103 news articles

Table 6. 5 Article Set Two Read by AudioBrowser

Purpose	Used by the participants during the final experiment session
Resource	▪ New York Times (November 2005), ▪ The Times Magazine Special Issue (April 2005), ▪ Listening comprehension test from the ETS TOEFL Test
Size of article set	Three basic categories that consist of a total of seven subsections and 50 articles

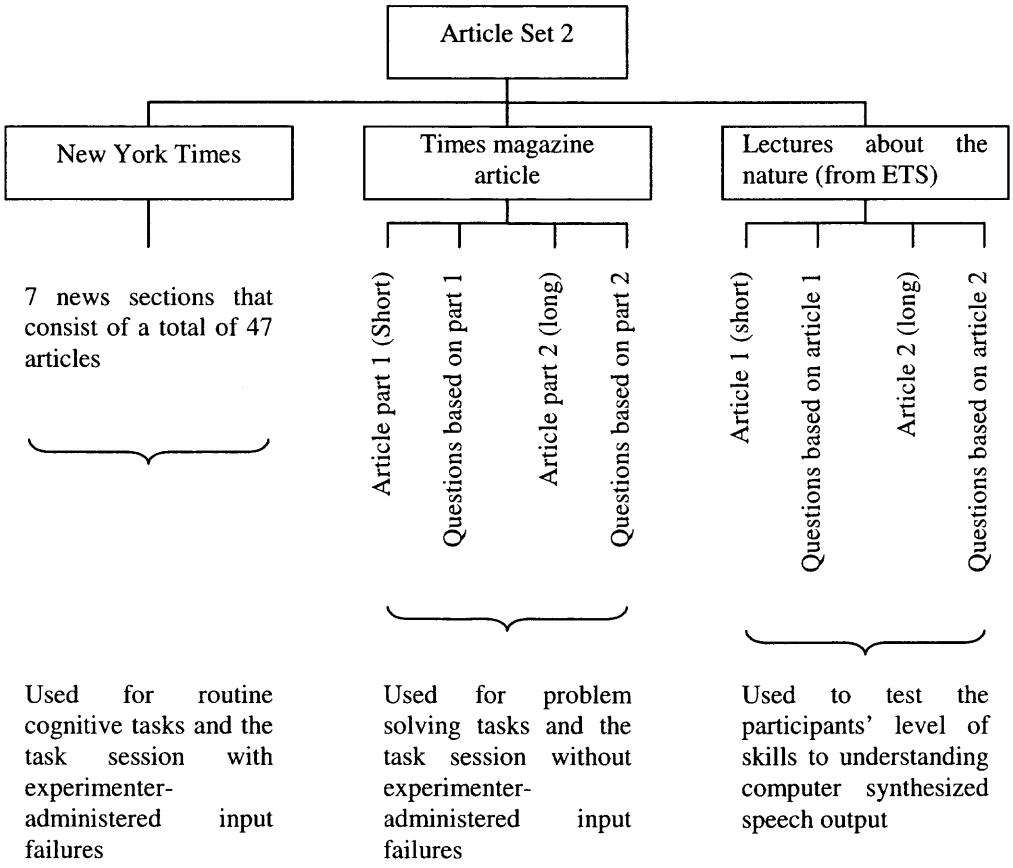


Figure 6. 3 Structure of Article Set 2 Read by AudioBrowser

6.4.4 Procedure Overview

Each subject participated in the experiment individually. The sessions for each subject spanned two days, with about two hours on each day. The following table shows the experiment procedure, estimated time of steps, documents / materials in each step, and data capture methods used. For the documents used during the experiment, please refer to Appendices A to K.

Table 6. 6 Procedure of the Controlled Experiment

Day	Step	Time	Documents / Materials	Data Capture
Day One	1. Thank-you to the subject; study and procedure introduction	5 minutes	Study introduction and procedure	
	2. Signing the consent form	10 minutes	Consent form	
	3. Pre-experiment interview about the subject's background	10 minutes	Background questionnaire	Questionnaire
	4. Tutorial on using speech and touch input	60 minutes	Training script, news articles (set 1)	
	5. Practice on speech and touch input	35 minutes	User tasks (set 1), news articles (set 1)	Videotaping
	Total time for Day One: 2 hours			
Day Two	6. Speech and touch input review	10 minutes	Summary from the speech and touch input tutorial	
	7. Warm-up practice session	15 minutes	User tasks (set 2), news articles (set 1)	Videotaping
	8. The experiment session:	65 minutes		
	8.1 Finding audio setting levels most comfortable to the subject		News articles (set 2)	Videotaping
	8.2 Testing the subject's ability to comprehend computer-synthesized speech output		Listening Comprehension articles from the TOEFL tests	Test answers
	8.3 Experiment task sessions		Experiment task script, news articles (set 2)	Videotaping
	9. Post-experiment interview	30 minutes	Post questionnaire	Videotaping & questionnaire
	Total time for Day Two: 2 hours			
Total time for the entire experiment with each subject: 4 hours				

The following paragraphs provide more details on each step in the experiment.

6.4.4.1 Steps on Day One.

Step 1: Study introduction. The introduction explained the purpose of the study to the participants, that is, to improve the design of the speech and touch input for visually impaired users. The introduction also explained the procedure and time needed.

Step 2: Signing the consent form. The consent form was approved by the NJIT Institutional Review Board (IRB). The consent form included a video and audio release agreement. The experimenter read the consent form for the subject and allowed enough time for questions. When all the questions were clarified, the experimenter helped the subject to find the location in the form for signature. The subject signed his/her name using regular handwriting on two paper copies of the consent form. The experimenter also signed on the copies. One copy was then kept by the experimenter. The other copy was given to the subject.

Step 3: Pre-experiment interview about the subject's background. The pre-experiment interview collected participants' background information including their age, educational background, visual impairment history, current vision, computer experience, and assistive technology support.

Step 4: Tutorial on using speech and touch input. For each user task, the subject was taught how to finish it using speech input and touch input. The tutorial combined learning and practicing. Exercise tasks were embedded in every paragraph of the tutorial. The experimenter read the tutorial and guided each subject's practice. Subjects were encouraged to ask questions at any time. Each tutorial lasted for about one hour.

The following paragraphs are examples from the tutorial script. The news articles used during the tutorial were from an old issue of New York Times.

> "...In front of you is a touchpad. You will use it to control the AudioBrowser system. [Let the subject feel the touchpad.] It is a rectangular device with an indented area at the center. The indented area can detect your touch and so is called the sensing area. The sensing area is divided into three tracks. [Guide the subject to feel the tracks.] The news sections and articles and system commands are mapped onto the tracks..."
>
> "...You can also browse the news categories using speech input. These are the commands you can use: ***next category***, ***next article***, and ***next item***. These commands let you go to the next news category or article available. ... Now please try these commands."
>
> "...You have learned how to use the touchpad and speech commands to browse new categories. Now you will finish some simple tasks using what you have learned. Please find the Sports Section using speech commands. Then find the Business Section using the touchpad."

Step 5: Practice on speech and touch input. A practice session of about 35 minutes followed the tutorial session. The subjects exercised speech input tasks, touch input tasks, mixed input modality tasks, and then had free choice of which input modality to use on a set of tasks. The purpose of these tasks was to allow the subjects not only to practice what they learned, but also to start forming their own mixed-modality input patterns based on their experience and preference.

Some examples of the practice tasks are shown below. The same set of news articles from the tutorial session was used.

- A speech task and a touch task: "Please use speech input to find an article titled *Johnson wins 200 meters semifinal* in the Sports section, and read the article sentence by sentence. Then pause after three sentences."

- "Please use the touchpad to find the third article in the Politics section, repeat the title of the article and spell the author's name."

- A mixed speech and touch input task: "Please use speech input to find an article in the Europe section about a political event that happened in Spain. Use the touchpad to decrease the reading speed by one level, and use the speech input to read the next sentence."

- A free choice of input modality task: "Please use the speech input and the touchpad input in any way you wish to do the following steps: find a news article interesting to you, spell the name of the author, increase the volume by two levels, and read a small part of the article."

The subjects were allowed to request the experimenter's help if they found it difficult to finish any task.

The experimenter administered a small amount of errors to allow the subjects to practice error correction.

6.4.4.2 Steps on Day Two.

Step 6: Speech and touch input review. On Day Two, the experimenter helped the subject review speech and touch input learned in the previous day by reading a summary of the system functions and speech and touch input commands.

Step 7: Warm-up task session. The warm-up session prepared the subjects for the experiment session. The types of tasks in this session were the same as those in the

practice session on Day One, including speech input tasks, touch input tasks, and free choice of input modality tasks.

Step 8: Experiment session. The experiment session was broken down into three phases: finding the subject's most comfortable speech settings, testing the subject's listening comprehension skill on computer-synthesized speech output, and administering the experiment tasks. Details of each phase are provided in Section 6.4.5, "Procedure of Experiment Session".

Step 9: Post-experiment questionnaire. After the experiment session, the subjects were interviewed about their experience and their opinion about using each input modality to finish specific tasks. The experimenter collected the subjects' opinions in forms of ratings and recorded their rationales for their ratings.

6.4.5 Procedure of the Experiment Session

The experiment session is one of the steps (i.e., Step 8 in Day Two) in the research.

Before the experiment session started, the subjects' freedom of choosing any input modality for any tasks was emphasized.

6.4.5.1 Step 1: Find subjects' most comfortable speech settings. Every subject was asked to go to the same article in the New York Times section, and adjust the audio settings while listening to the article. Each subject was asked to fine adjust speech and non-speech audio volumes, reading speed, and pitch, and select the voice with the pronunciation that was clearest. Each subject used as much time as needed to find the most comfortable audio settings.

Once a subject stopped adjusting audio settings, s/he was instructed not to change the settings any more until finishing Step 2 of the experiment session.

6.4.5.2 Step 2: Test subjects' ability to understand computer-synthesized speech output. A test examining the subjects' ability to understand synthesized speech was conducted in order to ensure that the ability to comprehend read text did not affect the subjects' performance during the experiment. During the test, the AudioBrowser system read two articles, a short one and a long one. Following each article, subjects were asked to answer comprehension questions about the article. Three questions followed the short article and four questions followed the longer article. The questions were multiple-choice questions each with A, B, C and D options. The questions required the subjects' understanding of the articles' contents and details.

The articles, questions and standard answers were from the Listening Comprehension Section of sample TOEFL tests (Test of English as a Foreign Language) published on ETS' web site (ETS, 2005).

Then with the audio settings selected during Step 1, each subject listened to each article once without interruption. Following each article, the questions were read to each subject as many times as needed before the subject gave his/her answer. To allow the subjects focusing on the listening comprehension task only, the experimenter used the touchpad to control the AudioBrowser to read the articles and the questions for the subjects.

The subjects' answers were recorded. A score was assigned to each subject based on the accuracy of the answers.

6.4.5.3 Step 3: Administer experiment conditions. The experiment conditions are shown in Table 6.3. The experimenter administered levels of error rates and types of cognitive tasks. The types of operators, navigation operators or non-navigation ones, were introduced by the experiment tasks – the tasks required the subjects to use both navigation and non-navigation operation to accomplish the tasks.

It has been explained that in the occasional input failures condition, the experimenter did not introduce any failure using the Wizard of Oz feature. The failures in this condition naturally occurred during the subjects' normal use. These failures included wrong user commands, uncorrected bugs in the AudioBrowser system, etc. In the frequent input failures condition the experimenter introduced a number of input failures simulating the failures that occurred in the natural context.

The data for each type of cognitive task and each type of operator were only collected from the occasional input failures condition, not the frequent input failures condition. There are two reasons for doing this: (1) The pilot study has shown that increased input failures change users' interaction patterns. Without artificially introduced errors, the occasional input failures condition has a higher chance to show users' normal interaction patterns. (2) This way of collecting data mitigates the situation of too many experiment conditions vs. small sample size.

6.4.5.4 Step 4: User interview and questionnaire. The interview and questionnaire were used to collect the subjects' opinions toward each type of input operator when completing different types of tasks. The interview and questionnaire were conducted between the occasional input failures condition and the frequent input failures condition.

The rationale was to get user opinions before they were changed by much higher input failure rates.

6.5 Summary

This chapter provided a detailed summary of the experiment design and experiment procedures used to test the research hypotheses generated in Chapter 5. In particular, it gave the rationale for the choices made and the controls embedded in the experiment design. The next chapter presents a detailed analysis of the data collected with these procedures and then discusses the findings uncovered by the research.

CHAPTER 7

OVERVIEW OF RESULTS FROM CONTROLLED EXPERIMENT

7.1 Results Overview

Findings from the controlled experiment are abundant. Chapters 7 through 12 are dedicated to present and discuss these findings. Among these chapters, Chapter 7 provides an overview of the results, and Chapter 12 provides a summary of the results. Chapters 8 through 11, each addresses one of the four research questions.

The most important findings presented in these chapters are:

- Most subjects selected multimodal input rather than unimodal input while both were available for free selection (Chapter 8).
- The subjects selected input modalities based on the type of input operation undertaken (Chapter 9).
- Input modality switching was not intensively used for error correction (Chapter 10).

An overview of the results for all research questions, as well as a summary of the subjects' background, is provided below.

1. The subjects' background (Chapter 7)
   In addition to satisfying the recruitment criteria, the subjects represented both a visually impaired population (some vision) and a blind population (no vision).

2. The subjects' ability to understand computer-synthesized speech (Chapter 7)
   Ten out of 19 subjects answered all listening comprehension questions correctly and obtained 7, the full score. Five subjects scored 6 out of 7. Three subjects scored 5 out of 7. And one subject scored 4. The correlation tests did not reveal any significant relationship between the subjects' listening comprehension scores and their choice of input modalities or their error correction behavior.

3. RQ1: users' choice between multimodal and unimodal input (Chapter 8)
   In the session with low error rates, five out of 19 subjects used unimodal input. Four of them used the touch input only and one used the speech input only. When error

rates were increased, only one out of 19 subjects used unimodal input. He stayed in the touch input mode.

4. RQ2: determinants of users' multimodal input patterns (Chapter 9)
   The type of input operation was found to have a significant influence on the subjects' choices of input modalities. Navigation operations were mostly performed using the touch input. Non-navigation operations were mostly finished using speech input.

   Cognitive task type affected the subjects' modality switching behavior. The subjects switched modalities significantly more for routine cognitive tasks than for problem solving tasks.

   The level of visual impairment was not found to influence a subject's choice of input modalities or modality switching behavior.

5. RQ3: effects of errors on users' usage of multimodal input (Chapter 10)
   Switching input modalities for error correction was not a common practice. The subjects were found to stay in the same input modality to correct errors.

   When error rates were increased, both modality switching for error correction and modality switching in general were increased. However, staying in the failing modality was still significantly more prevalent than switching the modality for error correction.

   The level of visual impairment was not influential to the subjects' modality switching behavior.

6. RQ4: common multimodal input patterns among sighted and visually impaired users (Chapter 11)
   Among the 33 sighted and visually impaired subjects, all but one visually impaired subject used multimodal input. The one who used unimodal input stayed in the touch input. The 32 sighted and visually impaired subjects who used multimodal input were found to choose input modalities based on the type of input operations undertaken. They seldom switched input modalities for error correction.

The organization of each of the chapters from Chapter 8 to Chapter 11 is as follows. The models and hypotheses of the research question being addressed in the chapter are restated. Data preparation and statistical method selection are then described. These are followed by the data analysis results. At last, discussions on the results are presented.

In the next section of this chapter, the subjects' background information and their scores in understanding computer synthesized speech are described.

7.2 Subjects' Background

Among the 20 recruited subjects, one male subject was not able to finish all the experiment tasks and hence, was excluded from the data analysis. The 19 subjects whose participation generated valid data for the research were between 20 and 60 years of age. The corrected vision in their better eye ranged from total blindness to being able to read magnified fonts. The numbers of the subjects in categories of vision, age and gender are shown in the Table 7.1.

Table 7. 1 Distribution of Subjects Based on Age, Gender and Level of Visual Impairment (I)

		Visual acuity in the better eye with correction							
		With working vision						With no working vision	
		20/80	20/100	20/200	20/400	20/800	20/1200	Some light perception	0
Age	20-29								Male: 1
			Female: 1		Female: 1				Female:2
	30-39						Male: 1	Male: 1	
				Female: 1					
	40-49				Male: 1	Male: 1			
		Female: 1							Female: 1
	50-60							Male: 1	Male: 3
					Female: 1				Female: 2

The subjects who only had light perception could perceive lights and shadows, but not any details of objects. So these subjects, together with those whose visual acuity was zero, were considered subjects with no working vision. The other subjects, to some

extent, depended on their visual acuity in their everyday life, and so belonged to the category that had working vision.

The distribution of subjects into the categories with and without working vision is shown in Table 7.2.

Except for one subject, all subjects' visual impairment resulted from one or more types of eye diseases, from Retinitis Pigmentosa (RP) to Retinopathy of Prematurity (ROP). One subject became visually impaired from a head injury.

Twelve out of 19 subjects were born with visual impairment. All subjects had visual impairment for five or more years.

Table 7. 2 Distribution of Subjects Based on Age, Gender and Level of Visual Impairment (II)

		Vision in the better eye with correction			
		With working vision		Without working vision	
Age	20-29	Male: 0		Male: 1	
		Female: 2		Female: 2	
	30-39	Male: 1		Male: 1	
		Female: 1		Female: 0	
	40-49	Male: 2		Male: 0	
		Female: 1		Female: 1	
	50-60	Male: 0		Male: 4	
		Female: 1		Female: 2	
Total		Male: 3	Sum: 8	Male: 6	Sum: 11
		Female: 5		Female: 5	

While the subjects had different education levels, all subjects had at least a high school diploma. Seven of them had bachelor degrees. Nine had already earned a college degree, or were in education programs working toward an advanced degree.

Eighteen out of 19 subjects had used computers for at least six years. One subject had been in assistive computer systems training classes for one year. All subjects used computers for at least an average of one hour per day. They used computers for a variety of activities from doing school and office work to entertainment and online shopping. The input methods of their assistive systems all included keyboard input, some had Braille input, but no speech or touchpad input. For all subjects but one, the output of their computer systems included speech output.

7.3 Subjects' Ability to Understand Synthesized Speech

In this experiment, it is important to address the issue on the participants' ability to understand computer-synthesized speech output. If their ability is not equal, then it is important to determine whether the differences in their ability could have influenced their multimodal interaction.

The average speed of the speech output the participants chose for the listening comprehension test was 225 words / minute, ranging from 160 to 320 words / minute. The distribution of listening speed choices is shown in Table 7.3.

Ten subjects obtained the full score in the text comprehension test. Five scored 6 out of 7. Three scored 5 out of 7. And one scored 4. The score distribution is shown in Table 7.4.

Table 7. 3 Listening Speed Selection

Listening speed (word / minute)	Number of participants
160	1
180	3
200	2
220	7
240	1
260	3
300	1
320	1

Table 7. 4 Listening Comprehension Scores

Listening comprehension score (Max. = 7)	Number of participants
7	10
6	5
5	3
4	1

In order to see whether the subjects' ability to understand computer-synthesized speech influenced their multimodal input choice, correlation tests were conducted. Kendall's tau-b, a non-parametric, rank-based correlation coefficient was used, because the listening comprehension scores were not normally distributed.

The results show that the correlation between the subjects' listening comprehension scores and the amount of speech inputs and touch inputs they used were not found to be correlated. The Kendall's tau-b coefficient was 0.266 but was not significant at $p < 0.05$. The scores were not correlated with the subjects' choice of error correction strategies either. A subject's error correction strategy was represented using

the percentage of times the user stayed in the same modality for error correction. The Kendall's tau-b coefficient was 0.094 and was not significant at $p < 0.05$.

The conclusion is that although differences exist in the subjects' ability to understand the computer-synthesized speech output used in this experiment, the differences was not found to influence the subjects' use of the multimodal input and hence, was not considered a factor that needs to be controlled.

CHAPTER 8

CHOICE BETWEEN MULTIMODAL AND UNIMODAL INPUT

8.1 Results

RQ1 asks "When interacting with a non-visual multimodal system, do visually impaired users use multimodal or unimodal input?"

To answer this question, descriptive statistics were looked at.

During the experiment sessions with visually impaired subjects, a total of 5519 input operators were performed by the 19 participants. Among these operators, 1683, or 30.49% were speech input, and 3836, or 69.51% were touch input. A total of 358 or 21.27% of speech input encountered errors. A total of 411 or 10.71% of touch input encountered errors. Out of the 5519 input operators, 755 or 13.68% involved switching from one input modality to a different one when the operator changed. All subjects, but one, switched input modalities during the complete experiment session.

Input operators executed by individual participants ranged from 177 to 462 and averaged 290.5 imput operators per subject. Speech input operators executed by individuals ranged from 0 to 269 and averaged 88.6speech input operators per subject. Touch input operators executed by individuals ranged from 11 to 462 and averaged 201.9 touch operators per subject.

Although most subjects switched input modalities, some subjects did not. This extreme preference over one input modality rather than the other was especially prominent when error rates were low. When error rates were increased, the participants' modality choices were distributed more evenly, on average.

In the experiment session with low error rates (i.e., error rates were between 3.35% and 24.45%), the participants executed a total of 4395 input operators, 1213, or 27.60% of which were speech input, and 3182, or 72.40% of which were touch input. 486 or 11.06% of the total number of 4395 input operators involved modality switches.

In the experiment condition with low error rates, five out of 19 subjects did not switch input modalities. Four out of the five subjects, who had no working vision, used touch input only; one of them, who had some working vision, used speech input only. In addition to their level of visual impairment, their choice of input modality did not appear to relate to any other background they had (including gender, age, years of computer experience, etc.).

When error rates were increased by 5.19% to 25.91% per subject, which resulted in 14.63% to 32.47% of failed input operators with each individual, the participants used speech and touch more evenly. In this condition with higher error rates, the participants executed a total of 1124 input operators, 470 or 41.81% of which were speech input, and 654 or 58.19% of which were touch input. The total number of operators involving modality switches was increased to 23.93%, or 269 out of 1124.

In the condition with high error rates, only one out of 19 participants chose unimodal interaction. This participant was one of the four who used touch input only when error rates were low.

Table 8.1 below summarizes these results.

Table 8. 1 Input Operators Executed during Experiment Sessions

	In the session with low error rates	In the session with high error rates	Sum
Total input operators	4395	1124	5519
Speech input (% of total input operators)	1213 (27.60%)	470 (41.81%)	1683 (30.49%)
Touch input (% of total input operators)	3182 (72.40%)	654 (58.19%)	3836 (69.51%)
Speech errors (% of speech input)	172 (14.18%)	186 (39.57%)	358 (21.27%)
Touch errors (% of touch input)	304 (9.55%)	107 (16.36%)	411 (10.71%)
Total switches (% of total input operators)	486 (11.06%)	269 (23.93%)	755 (13.68%)
No. of subjects not switching modality	5 out of 19	1 out of 19	--

8.2 Discussion

The important result obtained from the analysis above is that most subjects used multimodal rather than unimodal interaction when multimodal interaction was available. The secondary result is that, whether the error rates were low or high, the subjects mostly used more touchpad input than speech input.

The subjects' choice of multimodal, rather than unimodal interaction, is in accordance with the three-component working memory model (Baddeley, 1986; Quinn and McConnell, 1996). According to the model, human working memory has three components: the central executive, the phonological loop, and the visuo-spatial sketchpad. The phonological loop and the visuo-spatial sketchpad store distinct temporary information and process different tasks without interference against each other. The visually impaired subjects' phonological loop, or their verbal working

memory, stores and processes speech commands and synthesized speech output. Their visuo-spatial sketchpad, or the spatial working memory, processes very little to none of the visual information, but handles spatial information conveyed through the touchpad. The central executive supervises and coordinates tasks performed by the two subsystems by performing four functions: switching of retrieval plans, time sharing during dual-task performance, selective attention to certain stimuli while ignoring others, and temporary activation of long-term memory. It is believed that because of this working memory structure, the subjects were able to naturally divide spatial and verbal tasks between the two subsystems, and perform the tasks simultaneously.

In addition, the touchpad input allows exploration to find commands, which reduces the need to memorize commands. The speech input allows direct access to commands, which saves time since menu browsing is not needed. The two input modalities provided distinct advantages. The subjects switched between the modalities to get the most out of the multimodal interaction.

The fact that most subjects used more touchpad input than speech input could be explained as follows. Cognitive tasks during touchpad input involve processing spatial information conveyed through the touchpad, and processing computer speech output triggered by users' touch input. These tasks can be distributed to the spatial and the verbal subsystems for separate but simultaneous processes which take shorter times as compared to being processed linearly by a single subsystem. Cognitive tasks during speech input require more intensive use of the verbal subsystem to process both users' speech commands and computer-synthesized speech output. As such, speech input is more of a mono-subsystem task and requires a longer time for processing. The subjects

naturally used the touchpad input more often to offload some cognitive tasks to the spatial subsystem which otherwise would have to be processed by the verbal subsystem. This may also be the reason that when extreme preference is presented for one input modality over the other (i.e., when users do not switch modality at all), the preference is often given to the touchpad input – in the case of our study, four out of five subjects who did not switch input modality during the session with low error rates showed their loyalty to the touchpad input; and in the session with higher error rates, the only user who did not switch input modality stayed with the touchpad input, too.

In order to understand how the subjects used the two input modalities, and why a few subjects insisted on using unimodal input, further data analysis was conducted and is presented in the next two chapters.

CHAPTER 9

FACTORS DETERMINING MODALITY SELECTION

9.1 Overview

RQ2: Do any of the following factors have an impact on visually impaired users' multimodal input usage: type of input operator, level of visual impairment, and type of cognitive task?

In this research question, multimodal input usage refers to users' choice of input modalities and users' modality switches. Two models were constructed to investigate this question. The first model addresses the relationship between level of visual impairment, operator types, and users' choice of input modalities. The second model addresses the impact of cognitive task types on users' modality switching behavior.

9.2 Model 2.1: Effects of level of visual impairment and type of operator on choice of input modality

The independent and dependent variables of this model are:

- Independent variable (Between subject variable): level of visual impairment (with working vision vs. without working vision), and
- Independent variable (Within subject variable): type of operator (navigation operator vs. non-navigation operator)
- Dependent variable: Choice of input modality (speech input vs. touch input)

For calculation purposes, the dependent variable is represented using the ratio of speech inputs used over the total number of speech inputs and touch inputs used. Hence,

this ratio reflects both the amount of speech input and the amount of touch input. The formula is as follows:

Choice of input modality = number of speech inputs / (number of speech inputs + number of touchpad inputs)

The hypotheses for this model and the testing results are listed in Table 9.1.

Table 9.1 Hypotheses and Test Results for Model 2.1

	Hypothesis	Result
H2.1	Visually impaired users' choice of input modality is determined by the input operator type being executed, and users' level of visual impairment.	Partially supported
H2.1a	When performing navigation operations, users will use significantly more touchpad input and less speech input than when performing non-navigation operations.	Supported
H2.1b	Among visually impaired users, users with working vision will use the touchpad input significantly more than users with no working vision.	Rejected

Since five out of 19 subjects never switched input modality throughout the entire experiment session, to investigate whether these subjects' extreme interaction patterns have impacted the overall testing results, two sets of tests with and without the data from the five subjects have been conducted.

The following sections elaborate the hypotheses testing process.

9.2.1 Method Selection and Assumption Checking

Because there were two independent variables and one dependent variable, assumption checking was conducted to decide whether ANOVA or a non-parametric method was correct for hypotheses testing.

9.2.1.1 Assumption of Normal Distribution with Data from All Subjects. The subjects' choices of input modalities were checked in one batch. Since there were identical values in the data set, the Kolmogorov-Smirnov method was adopted for normality checking. The result showed that the significance value was less than .05. This meant that the null hypothesis that the data was normally distributed was rejected. The Normal Q-Q plot on the data showed a departure from the normal distribution. Table 9.2 and Figure 9.1 present the results of this test.

Table 9.2 Normality Test on Input Modality Choices

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Subjects' choice of input modality	0.1547	38	0.0222	0.8772	38	0.0006

a. Lilliefors Significance Correction

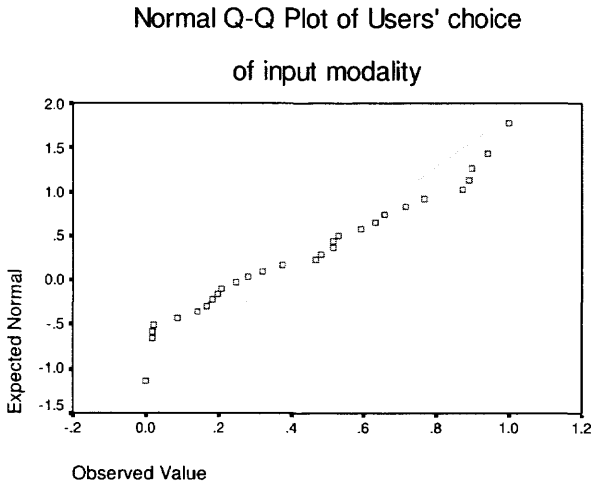


Figure 9.1 Normal Q-Q Plot of Input Modality Choices

To improve normality, the original data was transformed using the Arc-root transformation method. With this method, the arcsine values of the square roots of the original data were calculated. Normality was then checked based on the transformed data.

The normality testing result below shows that the transformed data was normally distributed. The significance value from the Kolmogorov-Smirnov test was larger than .05, therefore the null hypothesis that the data was normally distributed was not rejected. The Normal Q-Q plot shows an improved data distribution shape. These results are shown in Table 9.3 and Figure 9.2.

Table 9.3 Normality Test on Transformed Values of Input Modality Choices

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Subjects' choice of input modality (transformed)	0.1338	38	0.0836	0.9246	38	0.0136

a. Lilliefors Significance Correction

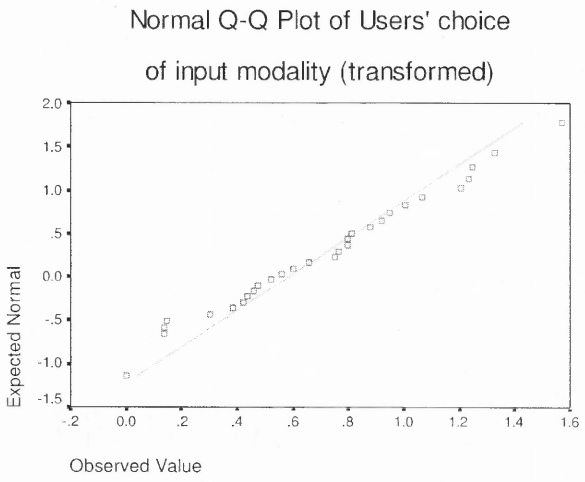


Figure 9.2 Normal Q-Q Plot of Transformed Values for Input Modality Choices

9.2.1.2 Assumption of Normal Distribution, Extreme Data Excluded: After excluding the data from the five subjects who never switched input modalities throughout the entire experiment session, the assumption of Normal Distribution was checked again using the data from the remaining 14 subjects. Since no identical values existed in the data sets this

time, the Shapiro-Wilks test was adopted. The testing result below shows that the significance value was larger than .05, hence the null hypothesis that the data was normally distributed was not rejected. No data transformation was needed. Table 9.4 and Figure 9.3 illustrate the result of this normality test

Table 9.4 Normality Test of Input Modality Choices (Extreme Data Excluded)

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Subjects' choice of input modality	0.1174	28	0.2 *	0.9378	28	0.0971

* This is a lower bound of the true significance

a. Lilliefors Significance Correction

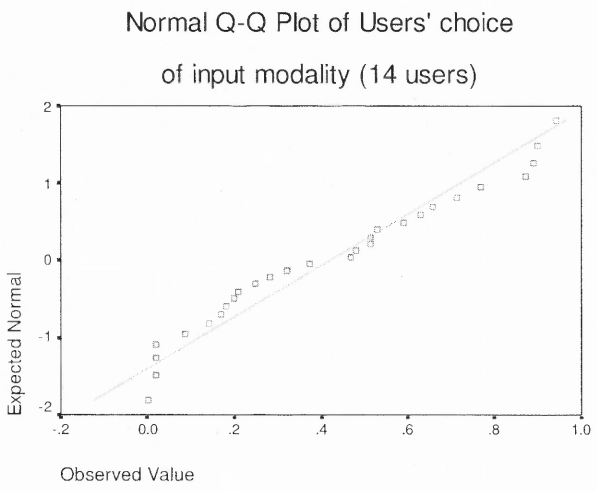


Figure 9.3 Normal Q-Q Plot of Input Modality Choices (Extreme Data Excluded)

9.2.1.3 Assumption of Homogeneity of Variance with Data from All Subjects. Levene's test of equal variance was conducted on the transformed data from all subjects. No significance less than .05 was found. Therefore the null hypothesis that the error variance of the dependent variable was equal across groups was not rejected. Table 9.5 presents the results from the homogeneity of variance test.

Table 9.5 Levene's Test of Equality of Error Variances on Modality Choices (All Subjects Included) [a]

	F	df1	df2	Sig.
Modality choice for navigation operators (transformed)	0.0076	1	17	0.9314
Modality choice for non-navigation operators (transformed)	2.2298	1	17	0.1537

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Vision; Within Subjects Design: Operator_type

9.2.1.4 Assumption of Homogeneity of Variance, Extreme Data Excluded. After excluding the data from the five subjects who did not switch input modalities during the experiment session, Levene's test of equal variance was used again to check the equal variance assumption using the data from the remaining 14 subjects. Again, all significance values were larger than .05. Therefore the null hypothesis that the error variance of the dependent variable was equal across groups was not rejected.

Table 9.6 Levene's Test of Equality of Error Variances on Modality Choices (Extreme Data Excluded) [a]

	F	df1	df2	Sig.
Modality choice for navigation operators (14 subjects)	0.5553	1	12	0.4705
Modality choice for non-navigation operators (14 subjects)	0.4727	1	12	0.5048

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Vision; Within Subjects Design: Operator_type

Based on the results of the assumption checks, ANOVA is the appropriate hypotheses testing method for both the data sets with and without extreme values.

9.2.2 Results

9.2.2.1 Results Based on Data from All Subjects. The between- and within- subjects variables formed four experiment conditions. The descriptive statistics and N of each experiment condition are shown in the following tables.

The Descriptive Statistics table shows that the subjects with working vision gave more speech input and less touch input than the subjects with no working vision – an average of 69.65% input operators executed by the subjects with working vision were speech input, while 50.04% input operators by the subjects with no working vision were given in speech.

The descriptive statistics also show that on average, the subjects used more touch input and less speech input for navigation operators than for non-navigation – on average, 43.73% of the navigation operators were performed using speech, while 72.87% of the non-navigation operators were performed using speech.

Table 9.7 Independent Variables and N for Model 2.1 (All Subjects Included)

Level of visual impairment (Between-subjects)	Type of operators (Within-subjects)	N
Subjects with working vision	Navigation operators	8
	Non-navigation operators	
Subjects with no working vision	Navigation operators	11
	Non-navigation operators	

Table 9.8 Descriptive Statistics for Model 2.1 (All Subjects Included)

	Level of visual impairment	Mean *	Std. Deviation	N
Modality choice for navigation operators (transformed)	With working vision	0.5333	0.4830	8
	With no working vision	0.3674	0.4167	11
	Total	0.4373	0.4408	19
Modality choice for non-navigation operators (transformed)	With working vision	0.8597	0.4130	8
	With no working vision	0.6334	0.5310	11
	Total	0.7287	0.4860	19

* Mean = the average percentage of user input given through the speech modality

ANOVA was used to test the effects of the within-subject and between-subject variables and their interaction effect using the data from all subjects. The results indicated that the Type of Input Operator had a significant effect on users' choice of input modality. The significance value was less than .001, with an observed power as high as .990. However, no significant effects were found from the level of visual impairment. Therefore, Hypothesis 2.1, that the level of visual impairment and the type of input operator being executed determine a user's choice of input modality, was partially supported. Hypothesis 2.1b that addresses how the level of visual impairment affects users' choice of input modality was rejected.

The analysis did not indicate any interaction effect between the level of visual impairment and the type of input operator.

Table 9.9 Test of Within Subjects Effects for Model 2.1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power (a)
Operator Type	0.8125	1	0.8125	20.7563	0.0003	20.7563	0.9900
Operator Type * Level of Visual Impairment	0.0084	1	0.0084	0.2157	0.6482	0.2157	0.0723
Error(Operator Type)	0.6655	17	0.0391				

a. Computed using alpha = .05

Table 9.10 Test of Between Subjects Effects for Model 2.1 (All Subjects Included)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power [a]
Intercept	13.2700	1	13.2700	33.5786	2.15E-05	33.5786	0.9998
Level of Visual Impairment	0.3563	1	0.3563	0.9016	0.3557	0.9016	0.1460
Error	6.7182	17	0.3952				

a. Computed using alpha = .05

The type of input operator has been demonstrated to affect users' input modality choice. To elaborate how it affected, paired T-Tests were conducted within the group of subjects that had working vision and the group of subjects that had no working vision. The results below indicate that for navigation operators the subjects with either visual impairment condition used significantly more touch input and less speech input than for non-navigation instructions. Therefore, Hypothesis 2.1a is supported.

Table 9.11 Effect of Operator Types on Input Modality Choices (Paired T-Test) (All Subjects Included)

	Subjects with working vision		Subjects with no working vision	
	Navigation operators	*Non-navigation operators*	*Navigation operators*	*Non-navigation operators*
Mean	0.28418	0.5562	0.2018	0.4364
Variance	0.1072	0.0932	0.0854	0.1423
Observations	8	8	11	11
Pearson Correlation	0.7495		0.7275	
Hypothesized Mean Difference	0		0	
df	7		10	
t Stat	-3.4209		-2.9995	
P(T<=t) one-tail	0.0056		0.0067	
t Critical one-tail	1.8946		1.8125	
P(T<=t) two-tail	0.0111		0.0134	
t Critical two-tail	2.3646		2.2281	

9.2.2.2 Results that Exclude Extreme Data. Because five out of 19 participants were found using mono-modality throughout the entire experiment session, there was a general interest in seeing whether those subjects' extreme input pattern had skewed the overall results. Therefore, a separate set of tests was conducted using data only from the remaining 14 subjects who switched input modalities during the experiment session.

The N's in the four experiment conditions formed by the between- and within-subjects variables, along with the descriptive statistics in the conditions are shown in the tables below.

The descriptive statistics obtained the same results as that before the extreme interaction data was removed, in that less speech operators and more touch operators were used for navigation operations than for non-navigation operations – an average of 48.12% of navigation operators were given in speech, while 87.67% of the non-navigation operators were given in speech.

Table 9.12 Independent Variables and N for Model 2.1 (Extreme Data Excluded)

Level of visual impairment (Between-subjects)	Type of operators (Within-subjects)	N
Subjects with working vision	Navigation operators	7
	Non-navigation operators	
Subjects with no working vision	Navigation operators	7
	Non-navigation operators	

Table 9.13 Descriptive Statistics for Model 2.1 (Extreme Data Excluded)

	Level of visual impairment	Mean *	Std. Deviation	N
Modality choice for navigation operators (Extreme values excluded)	With working vision	0.3851	0.2592	7
	With no working vision	0.5773	0.3847	7
	Total	0.4812	0.3306	14
Modality choice for non-navigation operators (Extreme values excluded)	With working vision	0.7581	0.3204	7
	With no working vision	0.9953	0.2231	7
	Total	0.8767	0.2924	14

* Mean = the average percentage of user input given through the speech modality

The descriptive statistic revealed a tendency opposite to that before the extreme interaction data was removed, in that the subjects with working vision used less speech input than the subjects with no working vision – an average of 57.16% input operators

were given in speech by the subjects with working vision, while 78.62% input operators were given in speech by the subjects with no working vision.

ANOVA was used to test the effects of the within-subject and the between-subject variables and their interaction. The results were the same as the results from the data from all subjects, but with a smaller significant value and a greater power: the Type of Input Operator had a significant effect on the subjects' choice of input modality (Sig. < .01, observed power = .9997).

However, no significant effects were found from the level of visual impairment, and the interaction between the level of visual impairment and the type of input operator.

Therefore, with data that excluded extreme input patterns, Hypothesis 2.1, that the level of visual impairment and the type of input operator being executed determine a user's choice of input modality, was again partially supported. Hypothesis 2.1b that addressed how the level of visual impairment affects users' choice of input modality was rejected.

Table 9.14 Test of Within Subjects Effects for Model 2.1 (Extreme Data Excluded)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power [a]
Operator Type	1.0948	1	1.0948	34.3663	7.6898E-05	34.3663	0.9997
Operator Type * Level of Visual Impairment	0.0035	1	0.0035	0.1112	0.7446	0.1112	0.0609
Error(Operator Type)	0.3823	12	0.0319				

a. Computed using alpha = .05

Table 9.15 Test of Between Subjects Effects for Model 2.1 (Extreme Data Excluded)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power [a]
Intercept	12.9078	1	12.9078	84.9491	8.581E-07	84.9491	1
Level of Visual Impairment	0.3227	1	0.3227	2.1238	0.1707	2.1238	0.2688
Error	1.8234	12	0.1519				

a. Computed using alpha = .05

Paired T-Tests were conducted to investigate how the Type of Input Operator affected Input Modality Choices. The one-tailed test was significant at the .05 level within both levels of visual impairment. Therefore, for navigation operators the subjects with either visual impairment condition used significantly more speech input and less touch input than for non-navigation instructions. Hypothesis 2.1a is therefore supported.

Table 9.16 Effect of Operator Types on Input Modality Choices (Paired T-Test) (Extreme Data Excluded)

	Subjects with working vision		Subjects with no working vision	
	Navigation operators	*Non-navigation operators*	*Navigation operators*	*Non-navigation operators*
Mean	0.1819	0.4928	0.3171	0.6858
Variance	0.0275	0.0713	0.0996	0.0376
Observations	7	7	7	7
Pearson Correlation	0.6083		0.6760	
Hypothesized Mean Difference	0		0	
df	6		6	
t Stat	-3.8809		-4.1782	
P(T<=t) one-tail	0.0041		0.0029	
t Critical one-tail	1.9432		1.9432	
P(T<=t) two-tail	0.0082		0.0058	
t Critical two-tail	2.4469		2.4469	

Therefore, the analysis from the data with and without the extreme user interaction data gave identical results.

9.2.2.3 Subjects' Subjective Ratings. The above statistics have provided evidence that the type of input operator to be executed has a significant effect on the input modality the user will choose to execute it. To understand how users would choose an input modality for each specific type of input operator, the subjects' subjective ratings were investigated at a detailed level.

Paired T-Tests were used to compare the subjects' ratings on speech and on touch. Because a Paired T-Test is robust against data departure from the normal distribution and the equal variances assumptions, assumption checking was not conducted.

Overall, the subjects did not find one input modality easier to learn than the other. The statistics shown in the following tables confirm this.

Table 9.17 Descriptive Statistics of Subjects' Overall Ratings on Speech and Touch

	Mean	N	Std. Deviation	Std. Error Mean
ease of learning of speech input (1 = very difficult; 5 = very easy)	4.1579	19	0.9582	0.2198
ease of learning of touch input (1 = very difficult; 5 = very easy)	4.1053	19	0.7375	0.1692

Table 9.18 Paired T-Test on Subjects' Overall Ratings

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
ease of learning of speech input - ease of learning of touch input	0.0526	1.3112	0.3008	-0.5794	0.6846	0.1750	18	0.8631

However, for each specific type of input task, the subjects rated on speech input differently than touch input.

9.2.2.3.1 Navigation tasks. In general, for five out of seven types of navigation tasks, the subjects rated the touchpad significantly easier to use than the speech input, and /or they were more likely to choose the touchpad than the speech input for the specific task. Their overall ratings for navigation operations are shown in the following tables:

Table 9.19 Overall Ratings on Speech and Touch for Navigation Operators

	Ratings on ease of use		Ratings on likelihood to use	
	Ratings on speech input	*Ratings on touch input*	*Ratings on speech input*	*Ratings on touch input*
Mean	1.7381	1.4444	2.1429	1.5079
Variance	0.2983	0.4871	0.4418	0.6944
Observations	18	18	18	18
Pearson Correlation	-0.1806		-0.3620	
Hypothesized Mean Difference	0		0	
df	17		17	
t Stat	1.2967		2.1727	
P(T<=t) one-tail	0.1060		0.0221	
t Critical one-tail	1.7396		1.7396	
P(T<=t) two-tail	0.2121		0.0442	
t Critical two-tail	2.1098		2.1098	

For each specific type of navigation operators, the subjects rated as follows:

(1) Browse news sections & article titles: To browse news sections and article titles of the same level on the information hierarchy, the user glides his finger across the touchpad track that holds the news sections / articles or says "next/previous category / article / item". For this operation the subjects didn't find one modality easier to use than the other, but would prefer to use touch than speech (Sig. < .05).

(2) Enter a news section: To enter a news-section (i.e., to browse information across different levels on the information hierarchy), a user clicks the right touchpad button or says "select / zoom in / read article". For this operation the subjects found the

speech input as easy to use as the touch input, but rated more likely to use touch than speech (Sig. = .053 when all subjects' data was counted, and Sig. < .05 when extreme interaction data was excluded).

(3) Exit a news section: To exist a news-section (i.e., to browse information across different levels on the information hierarchy), a user clicks the left touchpad button or says "exit" or "zoom out". For this operation the subjects' ratings were not significantly different between the to modalities in either ease of use or likelihood to use.

(4) When a user is in the "paragraph" mode, read the next paragraph: When the user is already in the "paragraph" mode, to read the next paragraph, the user either clicks the right touchpad button or says "next paragraph". No significant difference is found in the subjects' ratings on either ease of use or likelihood to use.

(5) Read five sentences continuously: To read five sentences continuously a user either touches the "sentence" unit on the touchpad and clicks the right button five times, or says "next sentence" five times. The subjects rated the touchpad input significantly easier to use than the speech input (Sig. < .05) and their likelihood to choose the touchpad input significantly higher than to choose the speech input (Sig. < .05).

(6) Browse available reading units: To browse the available reading units, a user glides his finger across the touchpad track that holds the reading units or says "next/previous reading unit". The subjects rated touch significantly easier to use (Sig. < .05) and having significantly higher likelihood for them to choose for this operation (Sig. < .05).

(7) Browse available audio settings: For this operation a user glides his finger on the touchpad unit that holds the audio settings or says "next/previous setting". The subjects rated that the speech input was easier to use for this task (Sig. < 0.1), and that it was more likely for them to choose the touchpad input than the speech input to do this task (Sig. < .05).

Table 9.20 Ratings on Each Type of Navigation Operators

		Average rating on ease of use*		Average rating on likelihood to use*		Sig. of Paired T-Tests (one-tailed) *	
		Speech	Touch	Speech	Touch	Ease of use**	Likelihood ***
1	browse news sections & article titles	1.8889 (1.8571)	1.7222 (1.5714)	2.6667 (2.6429)	1.7222 (1.6429)	0.3130 (0.1954)	0.0454 (0.0394)
2	enter a news section	1.3333 (1.2857)	1.3333 (1.1429)	1.8333 (1.7143)	1.2778 (1.1429)	0.5 (0.2173)	0.0531 (0.0357)
3	exit a news section	1.2778 (1.2143)	1.6111 (1.5714)	1.4444 (1.4286)	1.5556 (1.5)	0.1724 (0.1678)	0.3672 (0.4182)
4	In the "paragraph" mode, read the next paragraph	1.4444 (1.5)	1.2778 (1.1429)	1.6667 (1.5714)	1.4444 (1.3571)	0.2893 (0.1193)	0.2476 (0.2437)
5	read 5 sentences continuously	1.8889 (1.7857)	1.2778 (1.1429)	2.3889 (2.4286)	1.3333 (1.2143)	0.0469 (0.0347)	0.0031 (0.0007)
6	browse reading units	2.3889 (2.2857)	1.5 (1.3571)	2.6111 (2.7143)	1.6111 (1.4286)	0.0208 (0.0047)	0.0187 (0.0020)
7	browse audio settings	1.9444 (1.9286)	1.3889 (1.2857)	2.3889 (2.5)	1.6111 (1.4286)	0.0771 (0.0539)	0.0499 (0.0110)

* Results outside the parentheses are calculated based on data from all subjects; results in the parentheses excluded data from the subjects with extreme interaction patterns.

** Rating scale for ease of use: 1 = very easy to use; 5 = very difficult to use

*** Rating scale for likelihood to use: 1 = very likely to use; 5 = very unlikely to use

Sig. value with light underline: significant at the 0.1 level

Sig. value with heavy underline: significant at the 0.05 level

9.2.2.3.2 Non-navigation tasks. In general, for all five types of non-navigation tasks, the subjects rated the speech input significantly easier to use than the touchpad, and /or that they were more likely to choose the speech input than the touchpad for the specific task. Their overall ratings for non-navigation operations are shown in the following tables.

Table 9.21 Overall Ratings on Speech and Touch for Non-Navigation Operators

	Ratings on ease of use		Ratings on likelihood to use	
	Ratings on speech input	*Ratings on touch input*	*Ratings on speech input*	*Ratings on touch input*
Mean	1.1111	2.2778	1.2222	2.2889
Variance	0.1469	0.7112	0.1736	0.7210
Observations	18	18	18	18
Pearson Correlation	0.2628		0.2801	
Hypothesized Mean Difference	0		0	
df	17		17	
t Stat	-5.9664		-5.4227	
P(T<=t) one-tail	7.6607E-06		2.2853E-05	
t Critical one-tail	1.7396		1.7396	
P(T<=t) two-tail	1.5321E-05		4.5706E-05	
t Critical two-tail	2.1098		2.1098	

For each specific type of non-navigation operators, the subjects rated as follows.

(1) 1. Pause: To pause reading, the user either says "pause" or clicks the two touchpad buttons at the same time. The subjects' average rating values on speech and touch were not very different, but significance was found in the ease of use when all subjects' data was taken into account (that speech was significantly easier to use, with Sig. < .05) and in the likelihood to use when extreme interaction data was excluded (that they would significantly more likely to choose speech than touch, with Sig. < .05).

(2) 2. Resume: To resume reading, the user either says "resume" or touches the desired reading unit (i.e., word, sentence, or paragraph) on the touchpad and clicks the right button. The subjects preferred the speech input significantly more than the touchpad input (The significant values for both ease of use and likelihood are < .05).

(3) 3. Spell a word: To spell a word the user says "spell" or "spell word", or touches the word unit on the touchpad and clicks both touchpad buttons. Again, the subjects preferred speech significantly more than the touchpad (The significant values for both ease of use and likelihood are < .05).

(4) 4. Decrease reading speed: To decrease reading unit the user says "decrease speed" or touches the speed unit on the touchpad and clicks the left button. The subjects' preference on the speech input was prominent (The significant values for both ease of use and likelihood are < .05).

(5) 5. Repeat a sentence: To repeat a sentence, the user either says "repeat" or touches the sentence unit on the touchpad and clicks the left button to read the sentence again. Again, the subjects' preference on the speech input was prominent (The significant values for both ease of use and likelihood are < .05).

Table 9.22 Ratings on Each Type of Non-Navigation Operators

		Average rating on ease of use*		Average rating on likelihood to use*		Sig. of Paired T-Tests (one-tailed)*	
		Speech	Touch	Speech	Touch	Ease of use**	Likelihood ***
1	pause	1.1667 (1.2143)	1.6667 (1.4286)	1.5 (1.5714)	1.4444 (1.2143)	0.0230 (0.1361)	0.4208 (0.0481)
2	resume	1 (1)	3.2222 (3.0714)	1 (1)	3.0556 (3)	2.27E-07 (6.39E-06)	6.09E-06 (3.34E-05)
3	spell a word	1.1111 (1.0714)	2.0556 (2)	1.1667 (1.1429)	2.1667 (2.1429)	0.0015 (0.0053)	0.0016 (0.0009)
4	decrease reading speed	1.2222 (1.2143)	1.8889 (1.7857)	1.2778 (1.2857)	2.0556 (2)	0.0274 (0.0438)	0.0196 (0.0325)
5	repeat a sentence	1.0556 (1.0714)	2.5556 (2.3571)	1.1667 (1.2143)	2.7222 (2.5714)	2.94E-06 (2.95E-05)	1.83E-06 (2.11E-05)

* Results outside the parentheses are calculated based on data from all subjects; results in the parentheses excluded data from the subjects with extreme interaction patterns.

** Rating scale for ease of use: 1 = very easy to use; 5 = very difficult to use

*** Rating scale for likelihood to use: 1 = very likely to use; 5 = very unlikely to use

Sig. value with light underline: significant at the 0.1 level

Sig. value with heavy underline: significant at the 0.05 level

9.2.2.3.3 Special case. Setting the reading unit is a special case, because on the touchpad it is a navigation task (since the user glides on the touchpad track that holds reading units and stops his finger on the desired unit), but in speech it is a non-navigation instruction (since the user says "set to word /sentence /paragraph /article"). The subjects rated that they were more likely to use the touchpad input than the speech input to finish this task (Sig. < .05), reflecting that they preferred finding and executing the command through navigation rather than recalling and speaking it.

Table 9.23 Ratings on Setting Reading Unit

Operator	Type of operator	Average rating on ease of use *		Average rating on likelihood to use *		Sig. of Paired T-Tests (one-tailed) *	
		Speech	Touch	Speech	Touch	Ease of use **	Likelihood ***
set reading unit	Navigation on touchpad, Non-navigation in speech	1.8889 (1.8571)	1.6667 (1.5714)	2.2222 (2.0714)	1.4444 (1.2857)	0.2808 (0.2266)	0.0396 (0.0298)

* Results outside the parentheses are calculated based on data from all subjects; results in the parentheses excluded data from the subjects with extreme interaction patterns.

** Rating scale for ease of use: 1 = very easy to use; 5 = very difficult to use

*** Rating scale for likelihood to use: 1 = very likely to use; 5 = very unlikely to use

Sig. value with heavy underline: significant at the 0.05 level

The following charts summarize the subjects' ratings. With or without extreme interaction patterns, the subjects' average ratings show the same tendency on choices of input modalities.

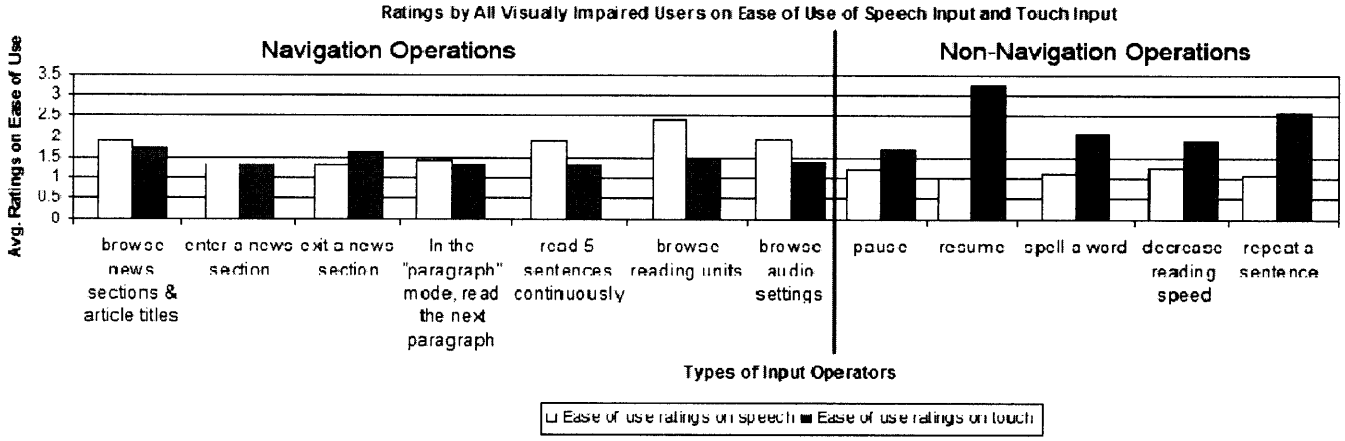


Figure 9.4 Ratings on Ease of Use on Speech Input and Touch Input (All Subjects Included)

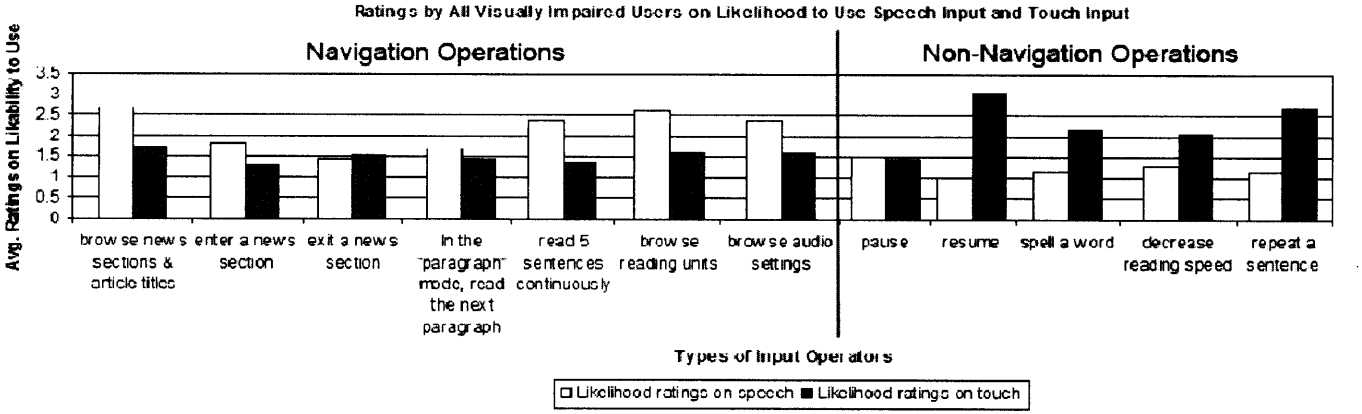


Figure 9.5 Ratings on Likelihood to Use on Speech Input and Touch Input (All Subjects Included)

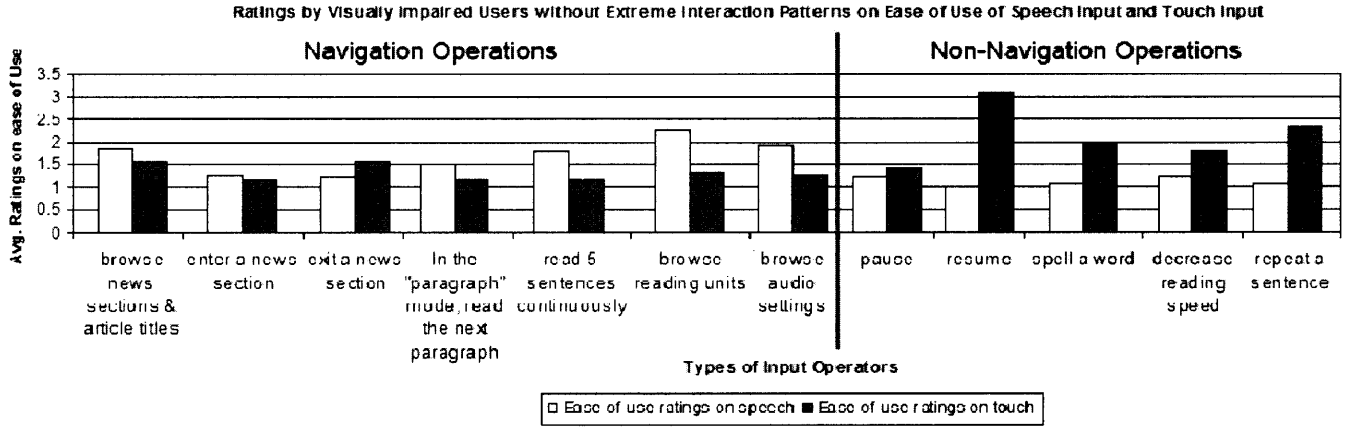


Figure 9.6 Ratings on Ease of Use on Speech Input and Touch Input (Extreme Data Excluded)

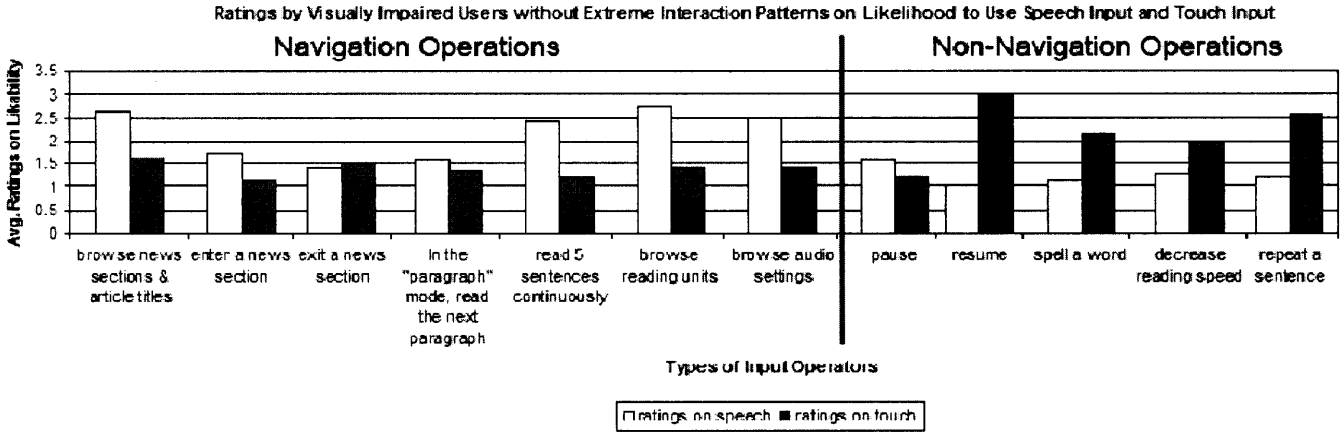


Figure 9.7 Ratings on Likelihood to Use on Speech Input and Touch Input (Extreme Data Excluded)

9.2.2.4 Extreme Interaction Patterns. To illustrate the data, the subjects' IDs are used.

In the experiment session with low error rates, five out of 19 visually impaired subjects used unimodal input. The subjects are labeled S4, S8, S14, S16 and S17. When error rates were increased by the experimenter, four of the five subjects switched input modalities, but one subject, S14, still insisted on unimodal input.

The five subjects' speech and touch input usage is illustrated in the following table.

Table 9.24 Modality Usage by Subjects Who Used Unimodal Input

	S8	**S14**	**S4**	**S16**	**S17**
Gender	Female	Female	Male	Male	Male
Level of visual impairment	With working vision	With NO working vision	With NO working vision	With NO working vision	With NO working vision
Input choice in the session with low error rates	Speech input only	Touch input only	Touch input only	Touch input only	Touch input only
Input choice in the session with high error rates	Multimodal input	Touch input only	Multimodal input	Multimodal input	Multimodal input
Error correction strategy in the session with high error rates	Multimodal error correction	Unimodal error correction	Unimodal error correction	Unimodal error correction	Multimodal error correction

The five subjects' ratings on speech input and touch input are illustrated in the following table. S4 did not participated in the rating, and so he is not listed in the table.

Table 9.25 Subjective Ratings by Subjects Who Used Unimodal Input

		S8	S14	S16	S17
Input choice	In the session with low error rates	Speech input only	Touch input only	Touch input only	Touch input only
	In the session with high error rates	Multimodal input	Touch input only	Multimodal input	Multimodal input
Average ratings for navigation tasks	Ease of use	Speech was easier	Touch was easier	Touch was easier	Speech was easier
	Likelihood to use	Speech was more likely to be selected	Touch was more likely to be selected	Touch was more likely to be selected	Touch was more likely to be selected
Average ratings for non-navigation tasks	Ease of use	Speech was easier	Speech was easier	Speech was easier	Speech was easier
	Likelihood to use	Speech was more likely to be selected	Speech was more likely to be selected	Speech was more likely to be selected	Speech was more likely to be selected
General ratings	Overall	Speech was rated better	Touch was rated better	No difference between speech and touch	Speech was rated better
	Ease of learning	Speech was easier to learn	Touch was easier to learn	Touch was easier to learn	Speech was easier to learn
	Ease of use	Speech was easier to use	Touch was easier to use	Speech was easier to use	Speech was easier to use
	Level of frustration	Speech was less frustrating	Touch was less frustrating	No difference between speech and touch	Speech was less frustrating

In general, during the session with low error rates, S8 used speech input only, whereas S4, S14, S16 and S17 used touch input only. During the session with high error rates, S14 insisted on touch input, while the other four subjects started to switch modalities. During the session with high error rates, S8 and S17 corrected errors by modality switching, but S4 and S16 used unimodal error correction only.

However, the subjects' ratings were not always consistent with their actual use of the input modalities.

S8, the subject who used speech input only during the session with low error rates, rated speech input consistently better than touch input. Even for performing navigation operations, S8 felt speech easier to use and as such, she was more likely to use speech input for navigation.

S14, the subject who insisted on touch input only during the sessions with both low error rates and high error rates, rated touch higher than speech most of times. However, for non-navigation tasks she rated speech easier than touch, and that she would be more likely to choose speech for non-navigation tasks.

S16 and S17, who used touch input only during the session with low error rates, rated speech better than touch for some questions (refer to Table 7.30).

The data above revealed individual differences in modality selection, as well as that individuals' modality selection was not necessarily conscious.

9.2.3 Discussion

The most important findings from Model 2.1 are that users choose input modality based on the type of operation undertaken, and that individual differences in modality switching should be expected, despite the modality choice – operation type dependence.

The level of visual impairment was not found to affect the subjects' modality choices. The reason is discussed in this section.

9.2.3.1 Input modality choice – input operator type dependence. The hypotheses testing results proved that subjects used significantly more touch input and less speech input for navigation operators than for non-navigation instructions. The subjects' subjective ratings revealed the reasons with more details. For most types of navigation

operations the subjects felt the touchpad input significantly easier to use and that to finish those operations they were more likely to choose the touchpad input. For most types of non-navigation instructions the subjects felt the speech input significantly easier to use and that they were more likely to choose the speech input to finish those operations.

The reasons for the subjects' preference over the touchpad input for navigation tasks could be that the touchpad input makes navigation operations easier by allowing the visually impaired subjects to mapping the information structure onto a physical space. This mapping helps the subjects to understand and explore the information structure by providing physical references.

Furthermore, based on Baddeley's working memory model (Baddeley, 1986), the working memory contains two subsystems for storage – phonological loop and visuo-spatial sketch pad. Although visually impaired users do not need to process visual information, their spatial sketch pad is used to process spatial information. Navigation operations often involve extensive speech output comprehension (e.g., consuming the meaning of sentences or paragraphs) that takes much use of the phonological loop in the working memory. The touchpad input, by allowing completion of navigational input using the spatial sketch pad in users' working memory, helps to avoid the chances of thresholds in the phonological loop and reduces the stress of the working memory.

The advantage of the touchpad is more prominent when continuous navigation is being executed. The statistical analysis indicated this in the following way: For doing a one-step navigation operation, no differences in the subjects' ratings on ease of use or likelihood to use were found between the two modalities, but when performing continuous navigation steps, the subjects rated that the touch input was significantly

easier to use than the speech input, and that their likelihood to use the touch input was significantly higher than the speech input. The underlying explanation used for this is that reading longer text results in a heavier workload in the phonological loop and hence, offloading some information processing tasks to the spatial sketch pad significantly reduced the burden on the subjects' working memory.

On the other hand, the visually impaired subjects preferred speech input for non-navigation instructions. This could be the result that (1) uttering a non-navigation command is quicker than searching for and executing the command on the touchpad, and (2) when the user is navigating the information space using the touchpad, giving a non-navigational command allows the user to keep tracking of his / her location on the touchpad (in the information space) and hence, is less interruptive of the navigation task.

Furthermore, many non-navigation instructions (e.g., change audio settings, set reading unit, spell a word, pause, etc.) are not followed by an extensive speech output comprehension task, i.e., the users do not need to understand the meaning of sentences or paragraphs following the instructions. Hence the phonological loop of the working memory has less working load compared to the workload following a navigational command. Completing the non-navigation operation on the touchpad to reduce the workload of the phonological loop is not necessary and takes longer than uttering a speech command. Therefore, users prefer to use speech input for non-navigation instruction.

One subject's comments during the interview complemented the point of view above: "[To change the reading speed,] when you are just starting to read an article, it's easy to do it using the touchpad, because you can find the command on the touchpad. But

when you are in an article already, it's definitely easier to say it than going through the settings [via the touchpad], because I can just automatically say 'increase speed'. It doesn't distract me from the reading"

9.2.3.2 Individual differences. When input errors were not manipulated by the experimenter, five out of 19 subjects exhibited extreme interaction patterns. S4, S14, S16 and S17, who had no working vision, used touchpad input only. S8, who had working vision, used the speech input only. After higher error rates were introduced, all of the five subjects but S14 used multimodal input. S14, who had no working vision, used touchpad input only.

The subjects explained their choices of input modalities during an interview between the session with low error rates and the session with high error rates. Their comments were therefore not influenced by the increased error rates in the second session.

S8, who used speech input only during the low-error rate session explained that "[Using the touchpad] is not as easy as saying it. You have to press more buttons and do more steps. When you verbally speak it, it will take you there in just one step."

S17's comments represented the opinions of the four subjects who stayed in touch mode during the low-error rate sessions: "[speech and touch] both have a proper place for use, if you said which one ... like I have to have one, then I probably will take the touchpad. Even though it might be more frustrating to find where it is, for me I think once you learn how to use it, you could use it. ... because for some reason to think [what the speech command is for] next sentence [is], compared to just do the next sentence [on the touchpad], is an extra brain step, which takes a little longer to me. But if you have no

use of hands, speech is excellent to read a newspaper." S14's comments complementary to S17's: "You know what it is about the speech ... If you don't remember the command, the touchpad can feed you back, but with speech, you cannot get anywhere if you don't give the right word". S16 explained his choice based on his experience with other systems: "I'm so used to using hand commands that it is always less energy than speaking".

It seems that the subjects' level of visual impairment determined their choice of input modality. However, hypothesis testing on the effect of visual impairment did not return any significant result. Moreover, the subjects' subjective ratings were not always consistent with their modality choices.

S8 and S14's overall ratings were consistent with their choice of input modality – S8 consistently rated speech input better than touch input, and S14 consistently rated touch input better than speech. S17, although used touch input only during the session with low error rates, consistently rated speech better than touch. S16, also used touch input only during the session with low error rates, however, rated speech easier to use, touch easier to learn, and no difference between the overall ratings and the level of frustration on speech and touch. (See Table 7.30)

For navigation operations, S8, S14 and S16's ratings on ease of learning, ease of use and likelihood to use were consistent with their usage of the modalities during the low error rate sessions – they rated the modality they each insisted on easier to learn, easier to use and that they were more likely to use it than the other modality that they did not choose. On the other hand, S17, who used touch only in the low error rate condition,

rated speech easier to learn and use than touch, but he admitted that he would be more likely to use touch for navigation tasks.

For non-navigation operations, all of the four subjects rated speech easier to use than touch, despite the fact that three of them only used the touch input.

The large difference in the subjects' choices of input modalities could be a result of the amount of navigation and non-navigation operations each individual executed, and the error rate each individual encountered.

During the task sessions with low error rates, S4, S14, and S17 performed more navigation operations than the other subjects. The navigation operations performed by S8 were close to the average of all subjects. In accordance, S4, S14, and S17 performed fewer non-navigation operations than all other subjects, while the amount of non-navigation operations performed by S8 was close to the average. Performing more navigation operations than other subjects might have encouraged S4, S14, and S17 to use more touchpad input than other subjects.

By using unimodal input, all of the five subjects achieved lower error rates than the other subjects. S14, the subject who insisted on unimodal input even when the error rate was increased, achieved the lowest error rate among all subjects during the session with low error rates. Under a lower error pressure than others, the five subjects were not as motivated to switch modalities for error correction as others.

When error rates were increased, the five subjects no longer had the lowest error rates. Four of them started to switch modalities. Two of the four started to use multimodal error correction. But the other two of the four insisted on unimodal error

correction instead of switching modalities to correct errors, despite having started to switch modalities for other tasks.

The explanation to the subjects' inconsistent ratings could be that their choices of input modalities were not always conscious. Users may have strong personal preferences that cause them to make their own modality choices and overwrite the common interaction patterns found in other users.

The above discussion implies, to some extent, that the distribution of the input operations used, in combination with error rates encountered, determined their individual preferences for one input modality over another. But the fact that their ratings were not consistent with their modality use implies that their choice of input modality was not entirely conscious. They naturally made a choice based on learned experience without much conscious thought.

9.2.3.3 Effect of level of visual impairment. Before the hypotheses testing it was believed that users with working vision would use the touchpad more than users with no vision. The rational was that people who had vision had an advantage in using the touchpad because they could rely on their vision to place their fingers on the desired location on the touchpad and hence assist their navigation of information through the use of a similar physical space. However, the descriptive statistics did not indicate this.

Table 9.26 Choice of Input Modality by Subjects with and without Working Vision

Data inclusion	Level of visual impairment	Choice of input modality
When all subjects' data was taken into account	Subjects with working vision	69.65% input operators were speech
	Subjects with no working vision	50.04% input operators were speech
When extreme user interaction patterns were excluded	Subjects with working vision	57.16% input operators were speech
	Subjects with no working vision	78.62% input operators were speech

Accordingly, the hypothesis addressing different choices of input modality by the subjects with working vision and without working vision was rejected by the data collected from the experiment.

Looking more closely, it was found that six out of eight subjects with working vision, who although confirmed that they somewhat depended on their vision in their everyday life, belonged to the legally blind category. (People whose best corrected vision is or less than 20/200 are legally blind.) Their vision might not be sufficient to help them in recognizing sensing tracks and buttons on the touchpad. Therefore there was no significant difference between them and the participants with no vision.

To evaluate whether the level of visual impairment really has no effect on users' input modality choices, the subjects with low vision but whose corrected vision is better than 20/200 should be recruited.

9.3 Model 2.2: Effects of cognitive task types on input modality switches

The independent variable of this model is cognitive task types (Routine Cognitive Tasks vs. Problem Solving Tasks). This is a within subject variable. The dependent variable is frequency of input modality switches.

The hypothesis for this model and the testing result are:

Table 9.27 Hypotheses and Test Results for Model 2.2

	Hypothesis	Result
H2.2	When performing routine cognitive tasks, users will switch input modality significantly more frequently than when performing problem solving tasks.	Supported

9.3.1 Method Selection and Assumption Checking

9.3.1.1 Adoption of Bootstrapping. In order to compare the subjects' modality switching behavior for routine cognitive tasks and problem solving tasks, data preparation and post-hoc control were conducted.

The pilot study with sighted subjects had revealed that the two most prominent factors that caused most input modality switches were the change of input operator types and the need for error correction. Therefore the major concerns of using the raw data for analysis of RQ2 without data preparation are as follows:

The first concern was that the level of error rate, in addition to the cognitive task type, might have skewed the subjects' modality switch patterns.

To check for the possible skew in data, a paired t-test was conducted to compare the error rates between the routine cognitive task session and the problem solving task

session. The result indicated no difference between the error rates in the two cognitive task sessions. Therefore, skew in data introduced by different error rates did not exist.

Table 9.28 Descriptive Statistics for Model 2.2

	Mean	N	Std. Deviation	Std. Error Mean
Error rate in routine cognitive task session	0.1244	19	0.0522	0.0120
Error rate in problem solving task session	0.1195	19	0.0749	0.0172

Table 9.29 Paired T Test Comparing Error Rates in Routine Cognitive Task Session and Problem Solving Task Session

Paired Differences					t	df	Sig. (2-tailed)
Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
			Lower	Upper			
0.0049	0.065	0.01483	-0.0263	0.0360	0.3298	18	0.7453

The second concern was that the transition between input-operator types (i.e., the transition between navigation operators and non-navigation operators), rather than the change in the cognitive task type, might have caused input modality switches.

To check whether this confound existed, a correlation analysis between input-operator type transition and the subjects' input modality switch was conducted. The correlation analysis was conducted both with data from all subjects and data only from the subjects without extreme interaction patterns.

In order to determine the correct correlation analysis method, data normality was checked within both the amount of modality switches and the amount of transitions between navigation and non-navigation operators. The normality testing results are shown in the following table.

Table 9.30 Normality Tests for Modality switches and Transitions between Operator Types

Data Inclusion	Variable	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
		Statistic	df	Sig.	Statistic	df	Sig.
Data from all subjects	TRANS	0.1302	38	0.1034	0.9623	38	0.2246
	SWITCHES	0.1664	38	0.0095	0.8968	38	0.0021
Data from the subjects without extreme interaction patterns	TRANS	0.1346	28	0.2*	0.9435	28	0.1354
	SWITCHES	0.1051	28	0.2*	0.9670	28	0.5023

TRANS = the amount of transitions between navigation and non-navigation operators

SWITCHES = the amount of input modality switches

* This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Since there was identical data within the amount of modality switches when all subjects' data is counted, the Kolmogorov-Smirnov test was used for the corresponding normality check. The Shapiro-Wilks test was used for all other normality checks.

The results indicated that the data from all subjects was not normally distributed, but that data from the subjects who did not present extreme interaction patterns was normally distributed. Therefore non-parametric correlation analysis (i.e., Spearman's rank correlation coefficient) was used for data from all subjects, and parametric correlation (i.e., Pearson's correlation) was calculated for data from the subjects without extreme interaction patterns.

The results of correlation analyses are shown in the following table.

Table 9.31 Correlation between Modality Switches and Operator Types Transitions

Data Inclusion			TRANS	SWITCHES
Data from all subjects	TRANS	Spearman's rho	1	0.3882*
		Sig. (2-tailed)	.	0.0160
		N	38	38
	SWITCHES	Spearman's rho	0.3882*	1
		Sig. (2-tailed)	0.0160	.
		N	38	38
Data from the subjects without extreme interaction patterns	TRANS	Pearson's r	1	0.4620*
		Sig. (2-tailed)	.	0.0133
		N	28	28
	SWITCHES	Pearson's r	0.4620*	1
		Sig. (2-tailed)	0.0133	.
		N	28	28

TRANS = the amount of transitions between navigation and non-navigation operators

SWITCHES = the amount of input modality switches

* Correlation is significant at the .05 level (2-tailed).

The above results clearly indicate that the input modality switches observed were partially caused by transitions between navigation and non-navigation operations.

Because of this confound, the raw data could not be used directly to test the relationship between cognitive task types and users' modality switches. Post-hoc control is necessary. The post-hoc control adopted was ***bootstrapping*** (Simon, 1997).

Bootstrapping is a resampling method with which a distribution is sampled multiple times to increase N. Bootstrapping is a popular testing mediation because it does not require data to be normally distributed and is effective with smaller sample sizes (N < 20) (Wikipedia, 2006).

Bootstrapping was conducted within the subjects' modality switches using the following steps:

(1) Each user's input operators were organized into four categories. The four categories are:

- Category 1: All operators in the routine cognitive task session that involve operator type transitions
- Category 2: All operators in the routine cognitive task session that do NOT involve operator type transitions
- Category 3: All operators in the problem solving task session that involve operator type transitions
- Category 4: All operators in the problem solving task session that do NOT involve operator type transitions

(2) Random selection was conducted repeatedly within each data category for each user:

- For each user, within each of the four categories above, a random selection of ten operators was repeated for ten times.

(3) Randomly selected data formed a pool of data for hypothesis testing:

- For each user within each of the four categories, the repeated random selection generated 100 input operators. Therefore for each user, 400 input operators were generated using bootstrapping. This equals to a total number of 3800 input operators from the routine cognitive task session and 3800 input operators from the problem solving task session.

Table 9.32 Distribution of Input Operators Generated Using Bootstrapping

Category		Number of Operators	Total Number of Operators
Operators from routine cognitive task session	Operators involving operator type transition	1900	3800
	Operators not involving operator type transition	1900	
Operators from problem solving task session	Operators involving operator type transition	1900	3800
	Operators not involving operator type transition	1900	

Therefore, by using the bootstrapping method, the effect of operator type transition on the subjects' input modality switch patterns was controlled.

The dependent variable was represented using the number of operators that involve modality switches.

A paired comparison between operators from each of the two cognitive task sessions was appropriate for the investigation of the proposed hypothesis.

Assumption checks were conducted to determine which paired comparison method should be adopted.

9.3.1.2 Assumption of Normal Distribution within Each Data Group. Since identical data existed in the data from all subjects, the Kolmogorov-Smirnov check was adopted. The results showed that data from all subjects could not be considered normally distributed. However, Paired T-Test was fairly robust to non-normality. So either a parametric or a non-parametric paired comparison could be used for the data set.

Table 9.33 Normality Test on Modality Switches (All Subjects Included)

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Amount of modality switches from the routine cognitive task session	0.2050	19	0.0346	0.8770	19	0.0191
Amount modality switches from the problem solving task session	0.1861	19	0.0823	0.8879	19	0.0295

a. Lilliefors Significance Correction

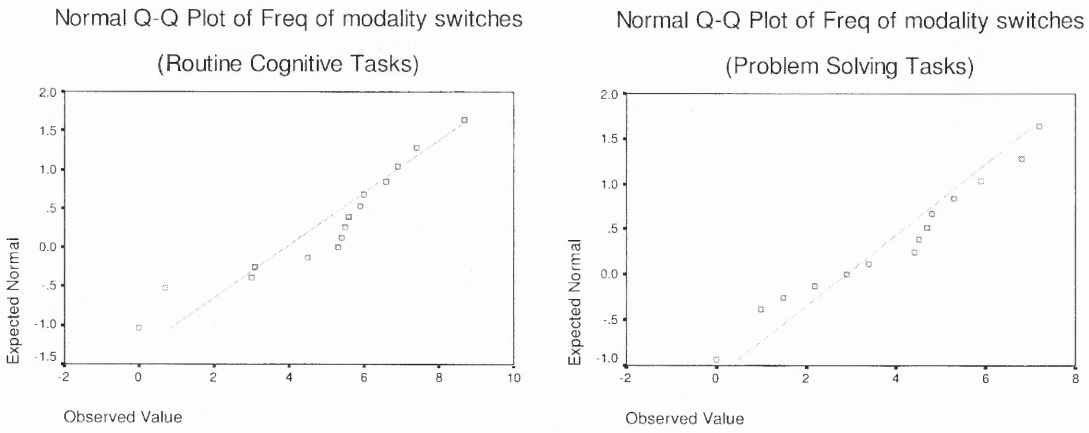


Figure 9.8 Normal Q-Q Plots of Modality Switches (All Subjects Included)

Since there was no identical data in the data from the subjects who did not present extreme interaction patterns, the Shapiro-Wilks test was adopted. The results showed that data from the subjects without extreme interaction patterns was normally distributed. Therefore a parametric paired comparison was appropriate for the proposed test.

Table 9.34 Normality Test on Modality Switches (Extreme Data Excluded)

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Amount of modality switches from the routine cognitive task session	0.2086	14	0.0999	0.9484	14	0.5366
Amount modality switches from the problem solving task session	0.1628	14	0.2 *	0.9690	14	0.8637

* This is a lower bound of the true significance.

a. Lilliefors Significance Correction

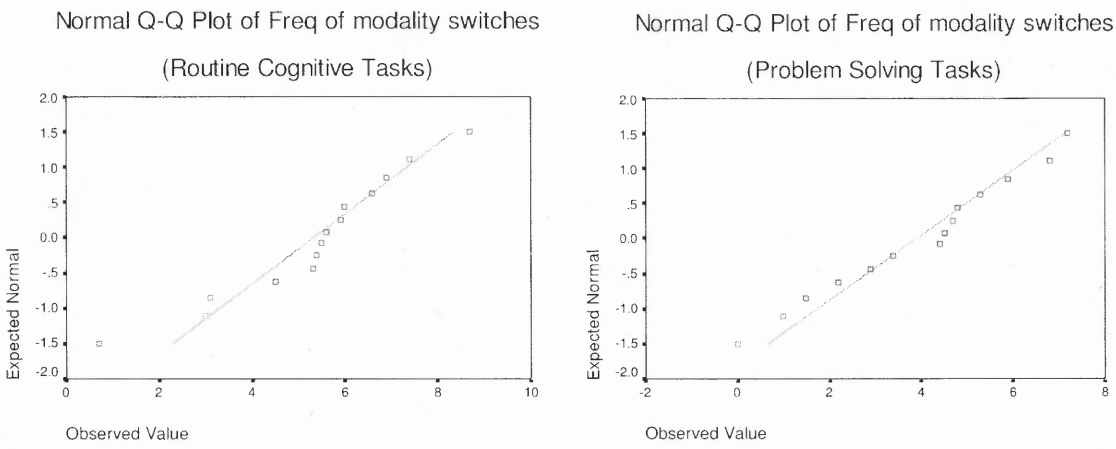


Figure 9.9 Normal Q-Q Plots of Modality Switches (Extreme Data Excluded)

9.3.1.3 Assumption of Normal Distribution within Paired Differences. Since there was no identical data in the paired difference between the two groups of data, the Shapiro-Wilks test was used for corresponding normality check. The result indicated a non-normal distribution. Again, since Paired T-Test was robust to non-normality, either a parametric or a non-parametric method could be used for the set of data from all subjects.

Table 9.35 Normality Test on Paired Differences (All Subjects Included)

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Paired difference	0.2092	19	0.0282	0.8756	19	0.0180

a. Lilliefors Significance Correction

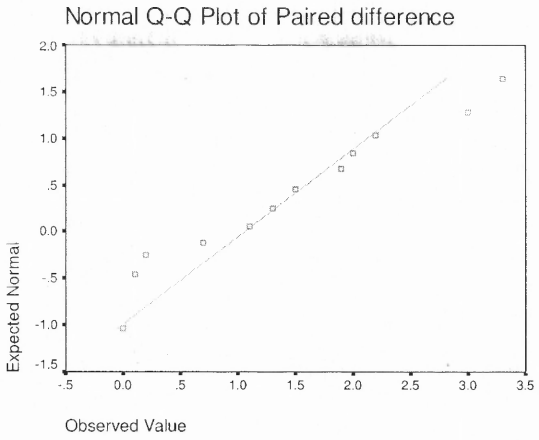


Figure 9.10 Normal Q-Q Plot of Paired Differences (All Subjects Included)

The Shapiro-Wilks test was used for the set of data from the subjects who did not present extreme interaction patterns. The result indicated a normal distribution within this data set. This confirmed the choice of a parametric method for hypothesis testing within this data set.

Table 9.36 Normality Test on Paired Differences (Extreme Data Excluded)

	Kolmogorov-Smirnov [a]			Shapiro-Wilks		
	Statistic	df	Sig.	Statistic	df	Sig.
Paired difference	0.1143	14	0.2 *	0.9485	14	0.5370

* This is a lower bound of the true significance.

a. Lilliefors Significance Correction

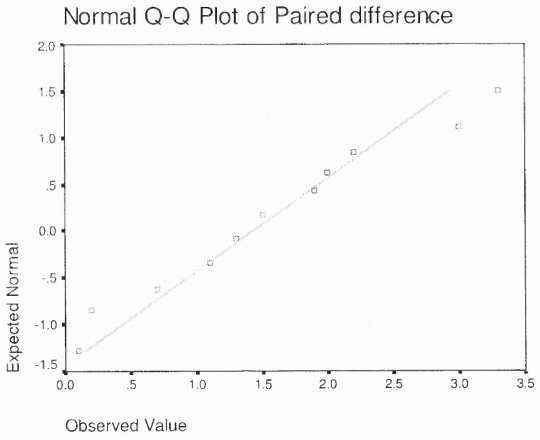


Figure 9.11 Normal Q-Q Plot of Paired Differences (Extreme Data Excluded)

9.3.2 Results

For data from all subjects, a Paired T-Test and its non-parametric version, a Wilcoxon Signed Ranks Test, returned the same results, which indicated that a significant difference existed between the amounts of switches from the two cognitive task sessions, and that the subjects switched input modalities significantly more frequently when performing routine cognitive tasks than performing problem solving tasks.

Table 9.37 Effect of Cognitive Task Type on Modality Switches (Paired T-Test) (All Subjects Included)

	Modality Switches from Routine Cognitive Tasks	Modality Switches from Problem Solving Tasks
Mean	3.9263	2.8737
Variance	8.7632	6.4932
Observations	19	19
Pearson Correlation	0.9360	
Hypothesized Mean Difference	0	
df	18	
t Stat	4.3050	
P(T<=t) one-tail	0.0002	
t Critical one-tail	1.7341	
P(T<=t) two-tail	0.0004	
t Critical two-tail	2.1009	

Table 9.38 Wilcoxon Signed Ranks for Modality Switches Based on Cognitive Task Type (All Subjects Included)

Pair		N	Mean Rank	Sum of Ranks
Modality switches from the routine cognitive task session - modality switches from the problem solving task session	Negative Ranks	14 [a]	7.5	105
	Positive Ranks	0 [b]	0	0
	Ties	5 [c]		
	Total	19		

a. Modality Switches from Routine Cognitive Tasks < Modality Switches from Problem Solving Tasks
b. Modality Switches from Routine Cognitive Tasks > Modality Switches from Problem Solving Tasks
c. Modality Switches from Routine Cognitive Tasks = Modality Switches from Problem Solving Tasks

Table 9.39 Effect of Cognitive Task Type on Modality Switches (Wilcoxon Signed Ranks Test) (All Subjects Included)

	Modality switches from the routine cognitive task session - modality switches from the problem solving task session
Z	-3.2982 [a]
Asymp. Sig. (2-tailed)	0.0010

Based on positive ranks.

For data from only the subjects who did not present extreme interaction patterns, a Paired T-Test was conducted. The result indicated that the subjects switched input modalities significantly more frequently when performing routine cognitive tasks than performing problem solving tasks.

Table 9.40 Effect of Cognitive Task Type on Modality Switches (Paired T-Test) (Extreme Data Excluded)

	Modality Switches for Routine Cognitive Tasks	Modality Switches for Problem Solving Tasks
Mean	5.3286	3.9
Variance	4.0868	4.68
Observations	14	14
Pearson Correlation	0.8886	
Hypothesized Mean Difference	0	
df	13	
t Stat	5.3600	
P(T<=t) one-tail	6.49E-05	
t Critical one-tail	1.7709	
P(T<=t) two-tail	0.0001	
t Critical two-tail	2.1604	

The conclusion from the results above is that input modality switches were significantly more frequent when the subjects were performing routine cognitive tasks than when performing problem solving tasks.

In addition to the above finding, the correlations between modality switches in the two cognitive task sessions were high (Pearson's Correlation = 0.9360 when all subjects' data is included; Pearson's Correlation = 0.8886 when only data from the subjects without extreme interaction patterns is included). The high correlation indicated a consistent modality switch pattern across different cognitive task types.

Table 9.41 Paired Sample Correlation between Modality Switches for Routine Cognitive and Problem Solving Tasks

Pair	Data Inclusion	N	Pearson's Correlation	Sig.
Modality switches in routine cognitive tasks - modality switches in problem solving tasks	Data from all subjects	19	0.9360	3.98E-09
	Data from only the subjects not presenting extreme interaction patterns	14	0.8886	2.1621E-05

9.3.3 Discussion

The input modality switch pattern discovered during the subjects' performance of routine cognitive tasks and problem solving tasks can be explained using Broadbent's bottleneck models (Broadbent, 1958), Kahneman's single resource pool models (Kahneman, 1973), and indications from previous research on how attention is divided between time-sharing tasks (Sweller, Chandler, Tierney & Cooper, 1990; Allport, Antonis & Reynolds, 1972; Shaffer, 1975; and Shiffrin, 1977).

According to bottleneck models, only a limited amount of information can be brought from the sensory register to the working memory for information processing. According to single resource pool models, attention is an information processing resource with limited capacity. Previous research indicates that how human allocates attention for different tasks competing for time and attentional resources depends on the degree to which one or more tasks can be performed automatically. When an automatic processing task is combined with any other more cognitively demanding task, more cognitive resources are available for the latter.

In the experiment sessions, a user's performance of a routine cognitive task was close to an automatic process due to the user's familiarity with the task. Performing a problem solving task, on the contrary, demanded more cognitive resources, normally involving extensive speech output processing. On the multimodal interface, switching modality demanded an increased level of cognitive resources which competed for time and cognitive resources with the performance of routine cognitive tasks and problem solving tasks. When routine cognitive tasks were processed in an automated fashion, more cognitive resources were available and, hence, encouraged modality switching. For problem solving tasks, since cognitive resources were limited, and more cognitive resources were demanded by the problem solving tasks, less resources were available and hence restricted modality switching.

The conclusion, therefore, can be made as follows: Cognitive task types have an impact on the frequency of switches between modalities – when performing routine cognitive tasks, users attempt to switch more frequently between speech and touch input than when performing problem solving tasks.

CHAPTER 10

EFFECTS OF ERRORS

10.1 Overview

RQ3: Will errors change users' multimodal interaction behavior?

This research question embraces three more detailed questions. For each of them, a quantitative analysis model was constructed. In addition, it was suspected that the subjects with working vision had been able to use their vision for information space path finding on the touchpad and, hence, could switch modalities more easily. Therefore, users' level of visual impairment was included as the second independent variable for two of the following models.

RQ3.1: Do users switch input modalities when correcting errors?

Model 3.1: Whether users are more likely to switch input modalities or use the same input modality to correct errors?

RQ3.2: Will level of error rates change users' error correction strategy?

Model 3.2: Whether users' modality-switching behavior for error correction is influenced by the level of error rates and users' level of visual impairment?

RQ3.3: Will level of error rates influence users' overall modality switching pattern?

Model 3.3: Whether users' modality-switching behavior in general, not just modality switches for error correction, is influenced by the level of error rates and users' level of visual impairment?

For Model 3.1, Hypothesis 3.1 was constructed. For Models 3.2 and 3.3, Hypotheses 3.2 (a & b) and 3.3 (a & b) were constructed. (See the table below.)

Models 3.2 and 3.3 have two different dependent variables but have the same independent variable. Therefore Models 3.2 and 3.3 were investigated together using multiple analyses of variance.

- The independent variables for Models 3.2 & 3.3 are:
    - Between subject variable: level of visual impairment (with working vision vs. without working vision)
    - Within subject variable: level of error rates (low error rates vs. high error rates)
- The dependent variables for Models 3.2 & 3.3 respectively are:
    - Input modality switches related to error correction
    - Total amount of modality switches
- In order to make data comparable among subjects, the raw data was processed using the following formulas:
    - Input modality switches related to error correction = The number of error correction operators that involved input modality switches / The total number of error correction operators
    - Total amount of modality switches in general = The total number of operators that involved input modality switches / The largest possible number of operators that could involve modality switches (i.e., the total number of operators -1)

    The hypotheses for RQ3 and the testing results are listed in Table 10.1.

Table 10.1 Hypotheses and Testing Results for RQ3

	Hypothesis	Result
H3.1	When errors occur, users will correct errors in the failing modality significantly more often than correcting them in another modality.	Supported
H3.2 & 3.3	Users' modality switches for error correction and modality-switching behavior in general are determined by the level of error rates and users' level of visual impairment.	Partially supported
H3.2 a	Users with working vision will switch input modalities more frequently for error correction than users with no working vision.	Rejected
H3.2 b	When error rate increases, users will switch input modality significantly more frequently for error correction.	Supported
H3.3 a	Users with working vision will switch input modalities more frequently in general than users with no working vision.	Rejected
H3.3 b	When error rate increases, users will switch input modality significantly more frequently in general.	Supported

10.2 Model 3.1: Modality switches for error correction

Model 3.1: Whether users are more likely to switch input modalities or use the same input modality to correct errors?

In order to provide analysis at a detailed level, three mean comparisons were conducted: a comparison between error corrections with modality switches and those without modality switches when error rates are low, the same comparison when error rates are high, and an overall comparison that embrace both error rate levels.

10.2.1 Method Selection

Paired T-Tests were appropriate for the mean comparisons. Because Paired T-Tests are robust against non-normal distribution, assumption check for normality was not conducted.

Since after error rates were increased, all subjects but one switched input modalities during task performance, the majority of the subjects did not present extreme interaction patterns. Therefore all subjects' data was included for hypothesis test.

10.2.2. Results

(1) Overall comparison:

An overall comparison was conducted between error corrections with modality switching and error corrections without modality switching. The results indicated that the subjects stayed in the same input modality significantly more often than switching to another input modality for error correction.

Table 10.2 Paired T-Test Comparing Overall Frequencies of Error Correction with and without Modality Switches

	Error correction using the same modality	Error correction by switching modality
Mean	30.8421	10.8421
Variance	110.8070	65.9181
Observations	19	19
Pearson Correlation	0.1551	
Hypothesized Mean Difference	0	
df	18	
t Stat	7.1128	
P(T<=t) one-tail	6.2648E-07	
t Critical one-tail	1.7341	
P(T<=t) two-tail	1.253E-06	
t Critical two-tail	2.1009	

(2) Comparison when error rates were low:

In the experiment session with low error rates, the subjects stayed in the same input modality significantly more often than switching to another input modality for error correction. The statistical results are displayed in the following table.

Table 10.3 Paired T-Test Comparing Frequencies of Error Correction with and without Modality Switches When Error Rates were Low

	Error correction using the same modality	Error correction by switching modality
Mean	20.3684	5.5263
Variance	99.5789	24.9298
Observations	19	19
Pearson Correlation	0.3460	
Hypothesized Mean Difference	0	
df	18	
t Stat	6.8184	
P(T<=t) one-tail	1.1022E-06	
t Critical one-tail	1.7341	
P(T<=t) two-tail	2.2044E-06	
t Critical two-tail	2.1009	

(3) Comparison when error rates were high:

In the experiment session with high error rates, the subjects stayed in the same input modality significantly more often than switching to another input modality for error correction. The statistical results are displayed in the following table.

Table 10.4 Paired T-Test Comparing Frequencies of Error Correction with and without Modality Switches When Error Rates were High

	Error correction using the same modality	Error correction by switching modality
Mean	10.4737	5.3158
Variance	20.4854	19.7836
Observations	19	19
Pearson Correlation	-0.2507	
Hypothesized Mean Difference	0	
df	18	
t Stat	3.1681	
P(T<=t) one-tail	0.0027	
t Critical one-tail	1.7341	
P(T<=t) two-tail	0.0053	
t Critical two-tail	2.1009	

10.2.3 Discussion

The important finding presented in this section is that, on an eyes-free interface, once an error occurs, switching modality is used significantly less by both sighted and visually impaired users than staying in the same input modality to correct that error.

This is significantly different from results of previous research on GUIs. Previous research on GUIs indicated that when one modality was failing, the allowance of modality switching made it easier to recover from the failure and, hence, was one of the prominent advantages provided by multimodal input.

The multiple resource pool theory (Navon and Gopher, 1979; Wickens, 1980, 1984 and 1992) can be used to explain this unwillingness to switch modalities for error correction, as well as the difference between the results of this research and previous research.

The multiple resource pool theory argues that instead of sharing a single pool of resource, there exist multiple pools of resources, each of which has its limited capacity and is related to specific skill. Multiple tasks can be performed at the same time as long as they require separate pools of resources.

When modality switching for error correction is to be performed, more than one task needs to be performed concurrently and therefore, competes for cognitive resources. The concurrent sub-tasks during error correction are: understanding input failure, finding solutions, and switching input modality. Unless modality switching is already a routine cognitive task that can be performed nearly in automation, modality switching will demand more cognitive resources from the two working memory subsystems, the phonological loop and the spatial sketchpad, which are already in heavy use. Through practice, the subjects learned their inability to processing these tasks simultaneously, which was the result of the bottleneck in their cognitive resources. Therefore the subjects avoid modality switching for error corrections unless the switch was a routine cognitive task.

On a graphical user interface where users can see the screen, the third sub-system in human working memory, the visuo-sketchpad can be used for task performance. The error correction sub-tasks can be divided among three working memory subsystems, the phonological loop, the visuo-sketchpad and the spatial sketchpad. The task load in each subsystem is therefore lower than when tasks are divided between two subsystems. Switching input modality, being an error correction method with a higher success rate, is then preferred by users.

This unwillingness to switch was presented in both the situation when error rates were low and the situation when error rates were high. But is a conclusion that allowing modality switch for the error-correction purpose is not desired by users at all correct? This question is answered by Model 3.2 by looking at whether the level of error rates and the level of visual impairment influence users' error-correction related modality switching pattern.

10.3 Models 3.2 & 3.3: Effects of Error Rates and Level of Visual Impairment on Modality Switching

Models 3.2 and 3.3 are:

- Model 3.2: Whether users' modality-switching behavior for error correction is influenced by the level of error rates and users' level of visual impairment?
- Model 3.3: Whether users' modality-switching behavior in general, not just modality switches for error correction, is influenced by the level of error rates and users' level of visual impairment?

10.3.1 Method Selection and Assumption Checking

Again, since after error rates were increased, all subjects but one switched input modalities during task performance, the majority of the subjects did not present extreme interaction patterns. Therefore all subjects' data was included for hypothesis test.

Because there were two dependent variables, both multivariate hypotheses (H 3.2 & H 3.3) and univariate hypotheses (H 3.2 a & b and H 3.3 a & b) were constructed. The multivariate hypothesis testing was used to reveal whether the independent variables had impacts on both dependent variables; while the univariate hypotheses testing was used to reveal the impact of the independent variables on each dependent variable.

There were two parametric multivariate methods available, MANOVA (Multivariate Analysis of Variance) and MANCOVA (Multivariate Analysis of Covariance). Because both independent variables were categorical rather than continuous, MANOVA, or its non-parametric alternative, would be the appropriate method to use. The following sections present the assumption checking to determine whether MANOVA or its non-parametric alternative should be used.

10.3.1.1 Assumption of Normal Distribution. Tests of normality were conducted within each of the four experiment conditions to decide whether a parametric statistical method should be adopted. Thee four conditions were:

- Condition 1: vision = with working vision, error rate = low error rate
- Condition 2: vision = with working vision, error rate = high error rate
- Condition 3: vision = with NO working vision, error rate = low error rate
- Condition 4: vision = with NO working vision, error rate = high error rate

Calculation based on two normality testing methods was conducted. Because the Shapiro-Wilks method has defects when identical values exist in the raw data, the Kolmogorov-Smirnov method was adopted instead.

Based on the Kolmogorov- Smirnov test, no results were significant at the .05 level, which indicated the data is normally distributed in all experiment conditions.

Table 10.5 Normality Tests on Modality Switches

Variables		Kolmogorov-Smirnov			Shapiro-Wilks		
		Statistic	df	Sig.	Statistic	df	Sig.
With working vision	Switches for error correction (when error rate is low)	0.2322	8	0.2	0.8917	8	0.2425
	Switches for error correction (when error rate is high)	0.1965	8	0.2	0.9313	8	0.5276
With NO working vision	Switches for error correction (when error rate is low)	0.1936	11	0.2	0.8797	11	0.1030
	Switches for error correction (when error rate is high)	0.2034	11	0.2	0.8821	11	0.1106
With working vision	Switches in general (when error rate is low)	0.1550	8	0.2	0.9505	8	0.7158
	Switches in general (when error rate is high)	0.1549	8	0.2	0.9192	8	0.4231
With NO working vision	Switches in general (when error rate is low)	0.2120	11	0.1799	0.8632	11	0.0634
	Switches in general (when error rate is high)	0.1567	11	0.2	0.9615	11	0.7897

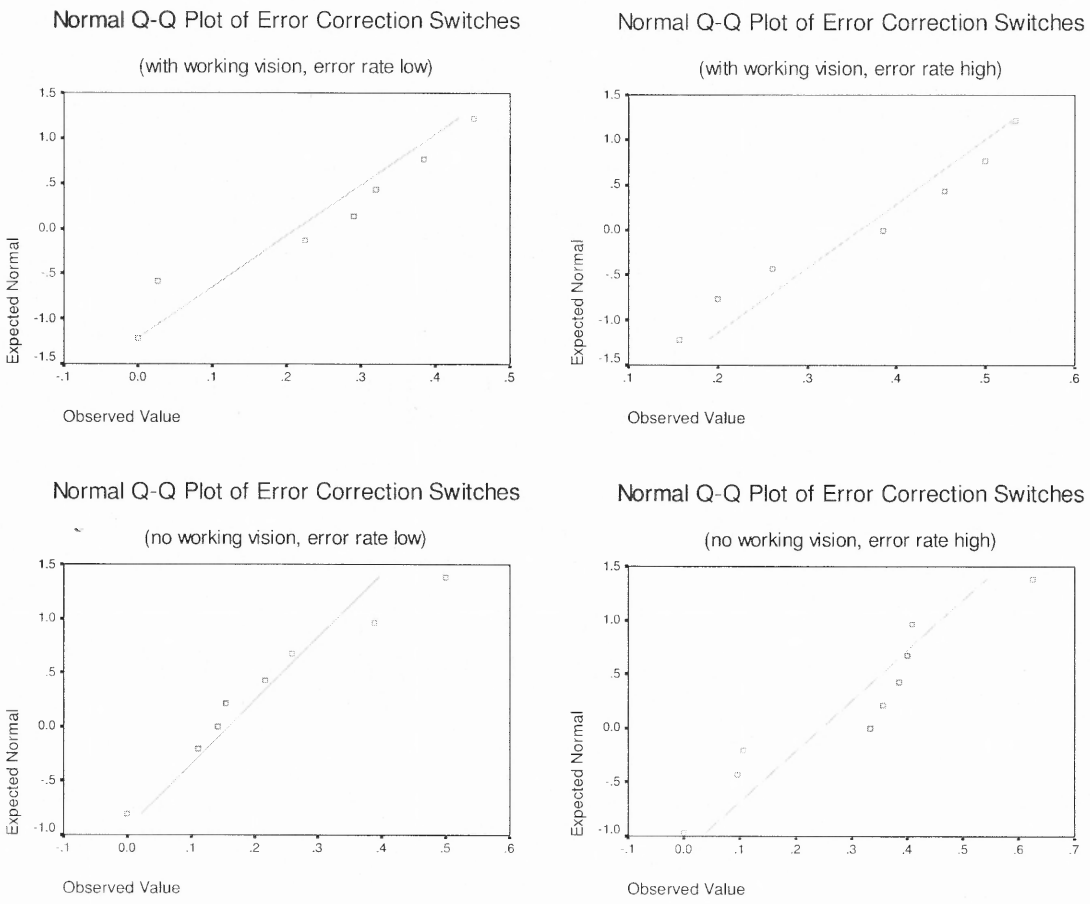


Figure 10.1 QQ Plots for Error Correction Related Input Modality Switches

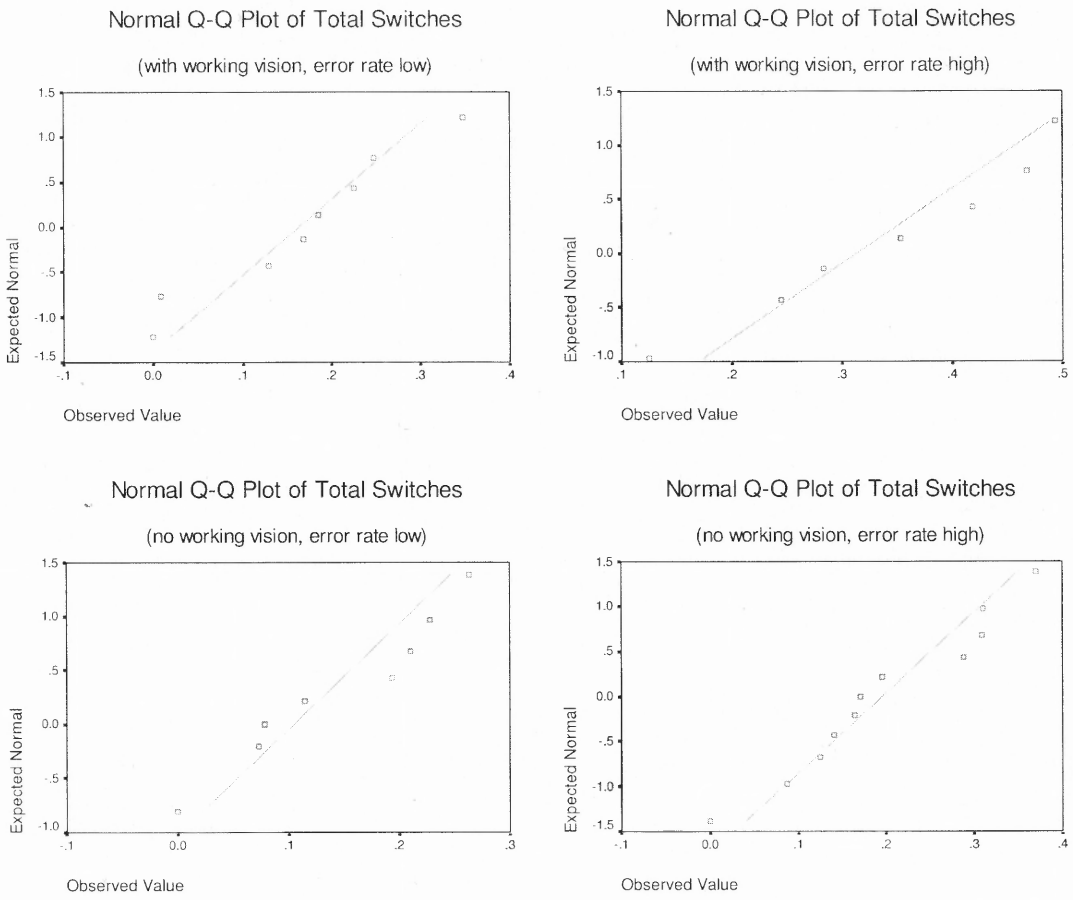


Figure 10.2 QQ Plots for General Modality Switches

10.3.1.2 Assumption of Homogeneity of Variance-Covariance Matrices. Box's M was used to test this.

Table 10.6 Box's Test of Equality of Covariance Matrices on Modality Switches

Box's M	9.2378
F	0.6714
df1	10
df2	1065.418
Sig.	0.7517
Design: Intercept+VISION; Within Subjects Design: ERROR RATES	

Box's M tested the null hypothesis that the observed covariance matrices of the dependent variables were equal across groups. Since the test result was not significant, the null hypothesis was not rejected. The assumption of homogeneity of variance-covariance matrices was supported.

10.3.1.3 Assumption of Homogeneity of Variance Levene's test of equal variance was conducted.

Table 10.7 Levene's Test of Equality of Error Variances on Modality Switches

	F	df1	df2	Sig.
Error correction switches (low error rates)	0.1689	1	17	0.6862
Error correction switches (high error rates)	4.1056	1	17	0.0587
Total switches (low error rates)	0.0006	1	17	0.9805
Total switches (high error rates)	1.0443	1	17	0.3212
Design: Intercept+VISION; Within Subjects Design: ERROR RATES				

This tested the null hypothesis that the error variance of the dependent variable was equal across groups. Since the test result was not significant, the null hypothesis was not rejected. The assumption was supported.

10.3.1.4 Assumption of Correlation between Dependent Variables. A two tailed correlation check was conducted between all observations of the dependent variables. It revealed that Pearson's r between the two dependent variables was .735, with sig. (2 tailed) < .001.

Table 10.8 Pearson Correlation between Modality Switches for Error Correction and in General

		Modality switches for error correction	Total modality switches
Modality switches for error correction	Pearson Correlation	1	0.7349*
	Sig. (2-tailed)	.	1.4906E-07
	Sum of Squares and Cross-products	1.2854	0.6866
	Covariance	0.0347	0.0186
	N	38	38
Total modality switches	Pearson Correlation	0.7349*	1
	Sig. (2-tailed)	1.4906E-07	.
	Sum of Squares and Cross-products	0.6866	0.6790
	Covariance	0.0186	0.0184
	N	38	38

* Correlation is significant at the 0.01 level (2-tailed).

The following line chart illustrates this correlation:

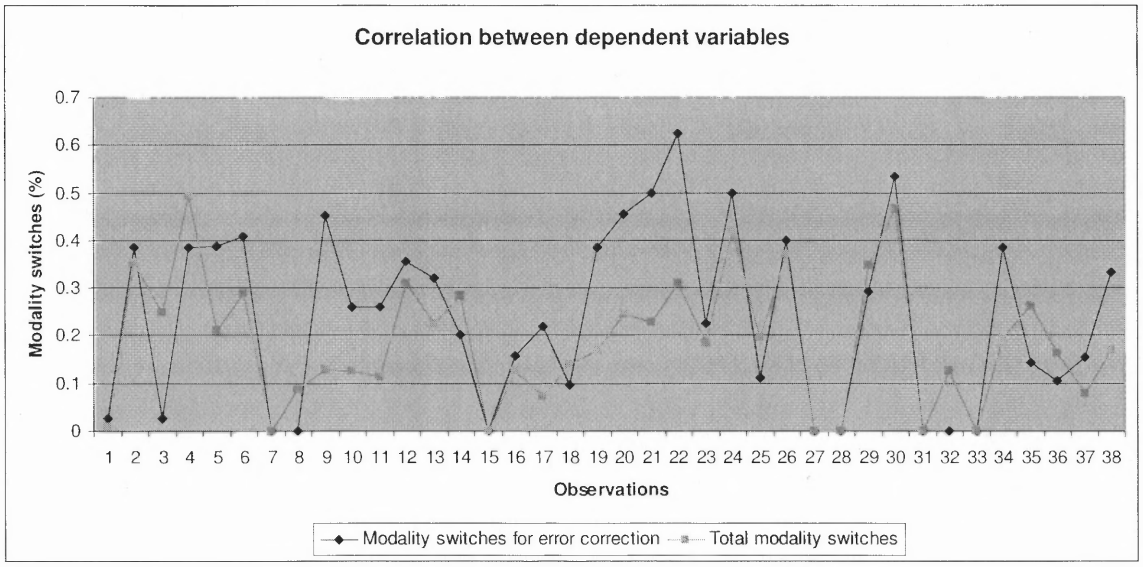


Figure 10.3 Correlation between Error Correction Related Modality Switches and General Modality Switches

10.3.2 Results

Both descriptive statistics and MANOVA (including multivariate analyses and univariate analyses) were used to test the hypotheses.

The descriptive statistics allowed comparisons of means in the experiment conditions. They revealed that in the condition with low error rates, both the average amount of error correction related input modality switches and the average amount of total modality switches were lower than the condition with high error rates. The statistics also showed that in all experiment conditions, on average, the subjects with working vision switched input modality more than the subjects with no working vision. When only looking at modality switches for error correction, on average, the subjects with working vision also switched more than the subjects with no working vision. The following tables show these statistics.

Table 10.9 Descriptive Statistics of Switches for Error Correction and in General

	VISION*	Mean	Std. Deviation	N
Switches for error correction (when error rates are low)	1	0.2157	0.1767	8
	2	0.1614	0.1692	11
	Total	0.1843	0.1697	19
Switches for error correction (when error rates are high)	1	0.3595	0.1394	8
	2	0.2463	0.2138	11
	Total	0.2940	0.1904	19
Switches in general (when error rates are low)	1	0.1640	0.1177	8
	2	0.1056	0.1025	11
	Total	0.1302	0.1100	19
Switches in general (when error rates are high)	1	0.3139	0.1442	8
	2	0.1963	0.1115	11
	Total	0.2458	0.1362	19

* Vision 1 = with working vision

* Vision 2 = with no working vision

The inferential statistics (MANOVA) allowed investigating of whether the above observations from the descriptive statistics could be generalized in the user population. The MANOVA tests included two sets of tests, the multivariate tests investigating the overall effects that the independent variables had on both dependent variables, and the univariate tests revealing the individual effects of the independent variables on each dependent variable.

A number of multivariate tests, which included Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root, were conducted. Among these statistics Wilks' Lambda is the choice for most researchers. The test of Wilks' Lambda, along with all other statistics, revealed that Error Rates had a significant effect on both dependent variables. The test was significant at .001 level, with an observed power of .996.

The results of the multivariate tests, hence, partially supported Hypotheses 3.2 & 3.3: A user's choice of error correction strategy, and the user's total modality switches in general are influenced by the level of error rates, but not the user's level of visual impairment.

Vision was not found significantly influential to the dependent variables. No interaction effect was discovered.

Table 10.10 Multivariate Tests for Models 3.2 & 3.3 ***

Effect	Variable	Method	Value	F **	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power *
Between Subjects	Intercept	Pillai's Trace	.796	31.173	2.000	16.000	.000	62.346	1.000
		Wilks' Lambda	.204	31.173	2.000	16.000	.000	62.346	1.000
		Hotelling's Trace	3.897	31.173	2.000	16.000	.000	62.346	1.000
		Roy's Largest Root	3.897	31.173	2.000	16.000	.000	62.346	1.000
	VISION	Pillai's Trace	.156	1.478	2.000	16.000	.258	2.956	.269
		Wilks' Lambda	.844	1.478	2.000	16.000	.258	2.956	.269
		Hotelling's Trace	.185	1.478	2.000	16.000	.258	2.956	.269
		Roy's Largest Root	.185	1.478	2.000	16.000	.258	2.956	.269
Within Subjects	ERROR RATE	Pillai's Trace	.650	14.882	2.000	16.000	.000	29.764	.996
		Wilks' Lambda	**.350**	**14.882**	**2.000**	**16.000**	**.000**	**29.764**	**.996**
		Hotelling's Trace	1.860	14.882	2.000	16.000	.000	29.764	.996
		Roy's Largest Root	1.860	14.882	2.000	16.000	.000	29.764	.996
	ERROR RATE * VISION	Pillai's Trace	.098	.873	2.000	16.000	.437	1.746	.174
		Wilks' Lambda	.902	.873	2.000	16.000	.437	1.746	.174
		Hotelling's Trace	.109	.873	2.000	16.000	.437	1.746	.174
		Roy's Largest Root	.109	.873	2.000	16.000	.437	1.746	.174

* Computed using alpha = .05

** Exact statistic

*** Design: Intercept+VISION+ ERROR RATE + VISION * ERROR RATE

Univariate Tests on the within-subjects variables revealed that the level of error rates had a significant effect on both the amount of modality switches for error correction (sig. = .013, with observed power of .747), and the amount of general modality switches (sig. < .001, with observed power of .998). No interaction effect was found.

Hypothesis 3.2 b was therefore supported: When users encounter higher error rates, they switch input modalities more often for error correction, as compared to the condition with lower error rates.

Hypothesis 3.3 b was also supported: In general, users switch input modalities more often when error rates increase.

Table 10.11 Tests of Within-Subjects Effects for Models 3.2 & 3.3

Source	Measure	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power [a]
ERROR RATE	Modality switches for error correction	.121	1	.121	7.756	.013	7.756	.747
	Total modality switches	.134	1	.134	27.140	.000	27.140	.998
ERROR RATE * VISION	Modality switches for error correction	8.009E-03	1	8.009E-03	.513	.484	.513	.104
	Total modality switches	8.083E-03	1	8.083E-03	1.636	.218	1.636	.227
Error(ERROR RATE)	Modality switches for error correction	.266	17	1.562E-02				
	Total modality switches	8.399E-02	17	4.940E-03				

a. Computed using alpha = .05

Univariate analyses on the between-subjects variable did not find any difference in the amount of error correction related modality switches between the subjects with or without working vision. The results revealed, however, that vision had a slight effect on modality switches in general (sig. = .094, with observed power of only .387).

Therefore, Hypothesis 3.3b, "In general, users with working vision switch input modalities more often than users without working vision", although not supported at the Sig. = .05 level, is supported at the Sig. = .1 level, but with a low power (Observed Power = .387). The lower power indicates that the conclusion that vision has an influence on the general users' amount of modality switches cannot be made unless an investigation with a larger sample size supports the same results.

The statistical results did not find the level of vision causing any difference in the amounts of error correction related modality switches among the subjects, and so H 3.2a was rejected.

Table 10.12 Average Amount of Modality Switches for Error Correction

	Error Rate =low	Error Rate =high
With working vision	0.2157	0.3595
Without working vision	0.1614	0.2463

Table 10.13 Average Amount of General Modality Switches

	Error Rate =low	Error Rate =high
With working vision	0.1640	0.3139
Without working vision	0.1056	0.1963

Table 10.14 Tests of Between-Subjects Effects for Models 3.2 & 3.3

Independent Var.	Dependent Var.	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power [a]
Intercept	Modality switches for error correction	2.237	1	2.237	45.689	.000	45.689	1.000
	Modality switches in general	1.408	1	1.408	61.672	.000	61.672	1.000
VISION	Modality switches for error correction	6.496E-02	1	6.496E-02	1.327	.265	1.327	.193
	Modality switches in general	7.170E-02	1	7.170E-02	3.140	.094	3.140	.387
Error	Modality switches for error correction	.832	17	4.897E-02				
	Total modality switches	.388	17	2.283E-02				

a. Computed using alpha = .05

10.3.3 Discussion

10.3.3.1 Effects of level of visual impairment in input modality switch patterns. The results could not prove the difference among the participants' level of visual impairment an impacting factor to the participants' modality switches either for error correction or in general. Although modality switches by the subjects with working vision was slightly significantly more than switches by the subjects with no working vision (Sig. = .094), the observed power (=.387) was too weak to conclude the acceptance of the hypothesis.

However, we cannot conclude that the level of visual impairment is not an impacting factor to users' modality switch patterns based on this study, because, as

explained in a previous section, most participants depending on their working vision to some extent in their life belonged to the legally blind category (i.e., six out of eight subjects with working vision were legally blind). The visual conditions between the group of the subjects with and without working vision might not have been big enough to make a difference in the comparison.

In order to make a conclusion, one more experiment is needed that uses participants with low vision who however do not belong to the legally blind category.

10.3.3.2 Effects of error rates in modality switches for error correction. Level of error rates has been proved to significantly affect users' modality switches for error correction and modality switches in general. When error rates increase, both types of modality switches are increased accordingly.

The reason of increased modality switches for error correction could be the following. The errors introduced using the Wizard of Oz method basically fell into the eight error categories discovered during the exploratory study with sighted subjects. In the experiment with visually impaired subjects, when error rates were increased, the subjects encountered more repetitive errors in each error category. It may be natural for the subjects to seek for alternative error correction methods when repetitive errors occur. Therefore they switched input modality to avoid same errors.

However it is noticeable that even the subjects' modality switches were increased when experiencing higher error rates, the subjects still preferred not switching than switching modalities for error correction. This was proved and discussed by the analysis of Model 3.1.

10.3.33 Effects of error rates in general modality switches. The reason that the subjects' modality switches in general were increased when error rates were higher could be the following. When the subjects switched more frequently for error correction, switching modality became a more familiar task. Being more familiar means becoming a more automatic process that requires less cognitive resources. Therefore the subjects were able to switch modalities more often in general.

CHAPTER 11

COMMON MULTIMODAL INTERACTION AMONG SIGHTED AND VISUALLY IMPAIRED USERS

11.1 Results

RQ4: Can we conclude any common or different patterns existing in sighted and visually impaired users' multimodal interaction?

In this research, the experiments with the sighted subjects and the visually impaired subjects were conducted separately following different procedures. Because of this reason, the comparison in multimodal interaction patterns between the two user groups was interpretation-based rather than hypothesis-testing-based.

There is also a general critique on a comparison between the two user groups, because so many differences exist, which just makes a comparison not possible.

However, the researcher still believes that a loose comparison based on interpretation of results from the two experiments will be helpful in understanding the multimodal usage differences and hence provide implications to designers who need to accommodate accessibility into their products.

The comparisons were conducted in three directions:

- Sighted and visually impaired subjects' adoption of multimodal input
- Sighted and visually impaired subjects' choice of input modalities
- Sighted and visually impaired subjects' error correction strategies

11.1.1 Sighted and Visually Impaired Subjects' Adoption of Multimodal Input

Given the same non-visual interface with integrated speech and touch input, and given that the overall error rates were at a similar level, the sighted subjects, on average, used slightly more speech input and less touch input than the visually impaired subjects.

Table 11.1 Overall Use of Input Modalities by Sighted and Visually Impaired Subjects

	Sighted subjects	Visual impaired subjects
N	14	19
Total No. of input operators	1642	5519
Total amount of failed operators	12.85%	13.68%
Total amount of operators involving modality switch	13.52%	13.68%
Total amount of speech operators	38.67%	30.49%
Total amount of touch operators	61.33%	69.51%
No. of subjects not switching input modality	0 out of 14	1 out of 19

11.1.2 Sighted and Visually Impaired Subjects' Choice of Input Modalities

11.1.2.1 Choice of input modality for each operator type. For navigation operators, both the sighted subjects and the visually impaired subjects used significantly more touch input than speech input. The amounts of speech and touch input the two user groups used were nearly identical (around 23% input were speech and around 77% input were touch).

For non-navigation operators, however, choices of input modalities were very different between the two subject groups. In general, the sighted subjects used significantly more speech input than touch input to accomplish non-navigation tasks (i.e., 61.5% of input was given using speech and 38.5% using touch). While the visually

impaired subjects did not show any significant modality choice pattern (i.e., 48.9% input was given using speech and 51.1% using touch).

Overall, the visually impaired subjects used nearly same amount of speech and touch input for non-navigation tasks. For some of the tasks, such as pause reading and changing audio settings, the visually impaired subjects used significantly more touch input. For some other tasks, such as resuming reading, they used significantly more speech input. This might have caused the overall touchpad usage by the visually impaired subjects higher than the overall touchpad usage by the sighted subjects.

Table 11.2 Use of Input Modalities for Each Operator Type by Sighted and Visually Impaired Subjects

Operator type		Sighted subjects (N=14)			Visually impaired subjects (N=19)		
		% of operators in speech	% of operators in touch	Significance (one-tailed paired t-test)	% of operators in speech	% of operators in touch	Significance (one-tailed paired t-test)
Navigation operators		**23.67%**	**76.33%**	**Sig. = .000**	**23.05%**	**76.95%**	**Sig. = .002**
1	Browse news sections & article titles on a single level	25.39%	74.61%	Sig. = .003	21.90%	78.10%	Sig. = .001
2	Go to a different information level (enter or exist a news section)	34.24%	65.76%	Sig. = .039	35.07%	64.93%	Sig. = .086
3	Proceed reading within text	18.67%	81.33%	Sig. = .000	23.49%	76.51%	Sig. = .004
Non-navigation instructions		**61.54%**	**38.46%**	**Sig. = .036**	**48.90%**	**51.10%**	**--**
1	Pause	57.70%	42.30%	--	33.80%	66.20%	Sig. = .041
2	Resume	94.62%	5.38%	Sig. = .000	96.21%	3.79%	Sig. = .002
3	Spell a word	53.50%	46.50%	--	48.41%	51.59%	--
4	Change settings' value	54.60%	45.40%	--	34.69%	65.31%	Sig. = .069
5	Repeat a sentence	100.00%	0.00%	Sig. = .011	48.28%	51.72%	--

11.1.2.2 Ratings on ease of use of input modalities for each operator type. When rating their overall preference on the speech and the touchpad input, neither the sighted subjects nor the visually impaired subjects presented a preference on one input modality over the other. However, when rating for each specific input operator type, for most operator types the subjects showed a preference on either speech or touch modality. The following table presents the details.

It should be noticed that the sighted subjects were asked to rate the ease of use and the likability on each input modality for each operator type, while the visually impaired subjects were asked to rate the ease of use and the likelihood to use each input modality for each operator type. Likability was described to subjects as "how much do you like to use [a specific input operator] to execute [a specific type of operator]". Likelihood was described as "how likely would you use [a specific input operator] to execute [a specific type of operator]". The two different ways of asking for subjects' ratings were intended to communicate the same meaning. This change of instrument was made during the controlled experiment because the experimenter found the second way of asking clearer.

To accomplish navigation operators, the sighted subjects felt touch input significantly easier to use than speech input, and liked the touch input significantly better. The visually impaired subjects, although did not feel touch input easier to use, would like to use touch rather than speech input to accomplish navigation tasks. Their preferences on touch input for navigation operators were the same.

However, differences existed when the two groups of subjects rated modalities for non-navigation tasks. Although they both expressed a preference on speech input, the

visually impaired subjects' preference on speech was stronger. The sighted subjects only rated speech significantly better for resuming reading. While the visually impaired subjects felt speech significantly easier to use than touch for all non-navigation operators. For three out of four non-navigation operators, the visually impaired subjects' ratings indicate that they would definitely choose speech rather than touch to finish the task. These results were interesting because the visually impaired subjects actually did not choose more speech than touch during their task completion, and for some non-navigation operations they actually chose more touch than speech input. The visually impaired subjects' subjective ratings were not consistent with their actually modality choices for non-navigation operators.

Table 11.3 Ratings on Input Modalities for Each Operator Type by Sighted and Visually Impaired Subjects

Operator type		Sighted subjects (N=14)			Visually impaired subjects (N=18)		
		Avg. ratings on speech input*	Avg. ratings on touch input*	Significance* (one-tailed paired t-test)	Avg. ratings on speech input**	Avg. ratings on touch input**	Significance** (one-tailed paired t-test)
Navigation operators		**3.224 (3.442)**	**1.806 (1.945)**	**Sig. = .001 (Sig. = .001)**	**1.738 (2.143)**	**1.444 (1.508)**	**-- (Sig. = .022)**
1	Browse news sections & article titles on a single level	3.702 (3.940)	1.417 (1.440)	Sig. = .000 (Sig. = .000)	1.889 (2.667)	1.722 (1.722)	-- (Sig. = .045)
2	Go to a different information level (enter or exist a news section)	2.155 (2.190)	1.714 (2)	-- (--)	1.306 (1.639)	1.472 (1.417)	-- (--)
3	Proceed reading within text	3.815 (4.196)	2.289 (2.400)	Sig. = .008 (Sig. = .006)	1.889 (2.389)	1.278 (1.333)	Sig. = .047 (Sig. = .003)
Non-navigation instructions		**1.863 (2.064)**	**2.523 (2.634)**	**Sig. = .033 (Sig. = .092)**	**1.111 (1.222)**	**2.278 (2.289)**	**Sig. = .000 (Sig. = .000)**
1	Pause	1.848 (1.981)	1.990 (2.183)	-- (--)	1.167 (1.500)	1.667 (1.444)	Sig. = .023 (--)
2	Resume	1.470 (1.663)	2.835 (3.186)	Sig. = .008 (Sig. = .007)	1 (1)	3.222 (3.056)	Sig. = .000 (Sig. = .000)
3	Spell a word	2 (2.214)	2.714 (2.571)	-- (--)	1.111 (1.167)	2.056 (2.167)	Sig. = .002 (Sig. = 002)
4	Change settings' value	1.929 (2.143)	2.524 (2.738)	-- (--)	1.222 (1.278)	1.889 (2.056)	Sig. = .055 (Sig. = .020)

* Results outside the parentheses are based on ratings on ease of use.

Results in the parentheses are based on ratings on likability.

Rating scale for ease of use: 1 = very easy to use; 5 = very difficult to use.

Rating scale for likability: 1= like to use a modality for a specific type of operator; 5 = dislike a modality for a specific type of operator.

** Results outside the parentheses are based on ratings on ease of use.

Results in the parentheses are based on ratings on likelihood to use.

Rating scale for ease of use: 1 = very easy to use; 5 = very difficult to use.

Rating scale for likelihood to use: 1= likely to use this modality for a specific operator type; 5 = not likely to use this modality for a specific operator type.

11.1.3 Sighted and Visually Impaired Subjects' Error Correction Strategies

For the exploratory study with the sighted subjects, the experimenter analyzed the types of errors occurred in speech and touch input. During the experiment with the visually impaired subjects, the experimenter introduced the same types of errors using the Wizard of Oz feature.

During the experiments, both the sighted subjects and the visually impaired subjects showed a strong preference on correcting errors using the same input modality rather than switching the modality.

Table 11.4 Adoption of Error Correction Strategies by Sighted and Visually Impaired Subjects

	Sighted subjects (N=14)			Visually impaired subjects (N=19)		
	Correction without modality switches	Correction with modality switches	Significance (one-tailed paired t-test)	Correction without modality switches	Correction with modality switches	Significance (one-tailed paired t-test)
Overall error correction	83.41%	16.59%	Sig. = .000	73.99%	26.01%	Sig. = .000
Correction of speech errors	83.23%	16.77%	Sig. = .000	65.13%	34.87%	Sig. = .022
Correction of touch errors	84.09%	15.91%	Sig. = .008	80.90%	19.10%	Sig. = .000

11.2 Discussion

11.2.1 Input Modality Choices by Sighted Subjects and Visually Impaired Subjects

The results with the visually impaired subjects have complied with the results with the sighted subjects in that the type of input operator is found significantly affecting subjects' input modality choices.

Before the comparison, it was believed that the sighted subjects would use speech input a little less than the visually impaired subjects. The rationale was that because the system output was speech, processing speech output and speech input competed for the same pool of resources in human attention and working memory. The visually impaired subjects were expected to be more skilled than the sighted subjects in processing speech input and output at the same time because their major information reception was through sounds.

However, the comparison showed opposite results. The visually impaired subjects actually used more touch input in general, and provided better subjective ratings on touch input than the sighted subjects for non-navigation tasks. There could be two reasons for the comparison results.

(1) Navigation in the information space was more cognitively demanding for most visually impaired subjects than for the sighted subjects. Navigating the information space requires tremendous cognitive resources for language processing. Therefore, the visually impaired subjects used touchpad more often to off load tasks processed in the phonological loop.

(2) Switching from touch input to speech input causes the lost of state on the touchpad. With the assistance of vision, the sighted subjects could get the state on the touchpad back quickly by landing their fingers on the previously touched location on the touchpad. The visually impaired subjects did not have this advantage brought by vision. Looking for the previous location on the touchpad required exploration through other locations on the touchpad, which some times caused accidental and unexpected command execution that made finding the previous location even more

difficult. Therefore, once "settled" on the touchpad for navigation operations, the visually impaired subjects were less willing to switch to speech input, despite the input operation became non-navigational. This explains why the visually impaired subjects actually used significantly more touch input than speech for some of non-navigation operations, but rated speech input significantly better than touch for all non-navigation operations.

11.2.2 Error Correction Strategies by Sighted Subjects and Visually Impaired Subjects

During error correction by both the sighted subjects and the visually impaired subjects, correcting errors using the same input modality was used significantly more frequently than correcting errors by switching the modality. This consistent behavior across different user groups indicates that, although a second modality provided an alternative way for error correction, subjects did not necessarily use the alternative way. The reason could be that switching modality is cognitively demanding. The subjects would rather repeat the failed input operation until it succeeds than making the efforts to seek for alternative error correction methods.

CHAPTER 12

SUMMARY OF RESULTS FROM CONTROLLED EXPERIMENT

In summary, the analysis of data from the controlled experiment revealed the following results. The results are organized by the research questions they address.

(1) RQ1: Do Users Use Multimodal Input

Most visually impaired subjects used multimodal rather than unimodal interaction when multimodal interaction is available. Individual differences existed – some subjects chose to stay in a single modality until error rates were increased. There was some suggestion that subjects with no vision tended to use only the touchpad but this was not supported by the data analysis. It may be possible that studies involving higher N might find this to be true for an important subset of subjects. This will require further investigation.

(2) RQ2: Multimodal Input Usage

Effect of Input Operation Type: Tests of hypotheses revealed that the type of input operation had significant impacts on users' multimodal input usage. For navigation operations, the subjects used significantly more touchpad input and less speech input than for non-navigation operations.

Effect of Cognitive Task Type: Tests of hypotheses revealed that the type of cognitive task had significant impacts on users' multimodal input usage. When performing routine cognitive tasks, the subjects switched input modalities significantly more frequently than when they performed problem solving tasks.

Effect of Level of Visual Impairment: The level of visual impairment was not found to significantly affect a users' multimodal input usage.

(3) RQ3: Multimodal Interaction Patterns When Errors are Present

RQ3.1: Do users switch input modalities for error correction?

Hypothesis testing revealed that, when errors occurred, visually impaired subjects continued to use the failing modality significantly more than switching to another input modality for error correction.

RQ3.2: Will the level of error rate and users' level of usable vision affect users' error correction strategy?

Hypothesis testing verified that error rates had a significant effect on users' error correction strategy. When error rates were increased, the subjects switched input modality more frequently to correct errors.

However, even when the subjects switched input modalities more frequently to cope with increased errors, the subjects in most occasions still stayed in the same modality rather than switching modalities for error correction.

The subjects' level of usable vision was not found to affect their error correction strategy.

RQ3.3: Will the level of error rate and users' level of usable vision influence users' modality switching behavior in general?

Hypothesis testing revealed that error rates had a significant effect on users' modality switching, in general, not merely when error correction was needed. When error rates were increased, the subjects switched input modality more frequently in general.

With the data collected in the experiment, the subjects' level of usable vision, again, was not found to affect their modality switching behavior in general.

(4) RQ4: Multimodal Input Usage by Sighted and Visually Impaired Users

Through comparison of results from the exploratory study with the sighted subjects and the controlled experiment with the visually impaired subjects, similar command interaction patterns were found in the subjects' choice of input modality based on input operation types, and the subjects' preferred strategy for error correction.

For both sighted and visually impaired subjects, input operation type significantly affected their modality choices. Navigation operations were usually performed using touchpad input, while non-navigation operations were frequently performed using speech input.

For both sighted and visually impaired subjects, correcting errors using the modality that was failing was more prevalent than correcting them in the other modality.

Table 12.1 summarizes the hypothesis testing results.

Chapters 7 to 12 listed all the findings from the experiment. The next and closing chapter attempts to put some meaning on these findings, in particular, discussing what the most significant results are in terms of (a) their likely impact on multimodal design and (b) their contribution to the theory of how people are likely to use multiple modalities.

Table 12.1 Summary of Results from Hypothesis Testing

		Hypothesis	Result
RQ2	H2.1a	When performing navigation operations, users will use significantly more touchpad input and less speech input than when performing non-navigation operations.	Supported
	H2.1b	Visually impaired users with working vision will use the touchpad input significantly more than users with no working vision.	Rejected
	H2.2:	When performing routine cognitive tasks, users will switch input modality significantly more frequently than when performing problem solving tasks.	Supported
RQ3	H3.1	When errors occur, users will correct errors in the failing modality significantly more often than correcting them in another modality.	Supported
	H3.2 a	Users with working vision will switch input modalities more frequently for error correction than users with no working vision.	Rejected
	H3.2 b	When error rate increases, users will switch input modality significantly more frequently for error correction.	Supported
	H3.3 a	Users with working vision will switch input modalities more frequently, in general, than users with no working vision.	Rejected
	H3.3 b	When error rate increases, users will switch input modality significantly more frequently, in general.	Supported

CHAPTER 13

CONCLUSION

The goal of this thesis is to understand how users coordinate hand and speech inputs to accomplish non-visual information browsing tasks in order to specify how similar systems should be designed.

In order to achieve this goal, the author conducted an exploratory study with sighted users, which refined the research questions and generated hypotheses about user multimodal hand and speech choices. Then a controlled experiment with visually impaired subjects was run to evaluate the hypotheses and provide answers to the research questions.

Abundant findings have been obtained. These findings, as well as the design implications, are presented in this chapter. In addition, contributions, limitations of the work, and future directions for the research are presented.

13.1 Summary of Findings

13.1.1 Multimodal Rather than Unimodal

All sighted subjects and most visually impaired subjects used multimodal rather than unimodal interactions when they had equal chances to choose between both.

The subjects' choice of multimodal input can be interpreted as resulting from the distinct advantages provided by each modality. The touchpad input reduces memorization load by allowing command search through menu browsing. The speech input provides direct access to commands and saves time by avoiding menu browsing, much in the way

macro commands work in text editing. The subjects naturally switched input modalities based on their needs. If they were close to their menu item, they were likely to use touch to advance to it. If they were not close, they were likely to use speech and to eventually memorize the speech command through frequent usage.

Moreover, from the psychological point of view, allocating tasks into different modalities allows concurrent task processing in working memory. According to Baddeley and Hitch's working memory model (Baddeley and Hitch, 1974; Baddeley, 2000), tasks on the touchpad are processed in the visuo-spatial sketchpad, and speaking and listening tasks are processed in the phonological loop. Using the touchpad for some input tasks offloads the burden that would otherwise be carried in the phonological loop, because both speech inputs given by the user and speech outputs given by the computer are processed in the phonological loop. By using the touchpad, tasks processed in the visuo-spatial sketchpad and the phonological loop are balanced and processed concurrently. The central executive then works to integrate information in the two subsystems.

In summary, it might be the above two reasons, stated from different perspectives, that have led to users' natural choice of multimodal, rather than unimodal interaction.

13.1.2 Modality Choice Based on Input Operation Type

More than choosing multimodal interaction, the subjects seemed to choose certain modalities for specific user input commands. For information space navigation operations, the subjects used significantly more touch input than speech input. For non-navigation operations, the subjects used significantly more speech input than touch input.

This modality choice – command type dependence, again, can be explained from two perspectives.

13.1.2.1 Explanation using the least effort point of view. From the psychological point of view, users choose input modalities that lead to least cognitive effort.

Navigation operations (e.g., going to the next paragraph) are often followed by intensive listening comprehension tasks, and consequently, intensive use of the phonological loop. Performing input tasks on the touchpad avoids increased workload in the phonological loop. Therefore, the touchpad is chosen for navigational inputs.

On the other hand, most non-navigation commands (e.g., reducing the reading speed) do not lead to intensive listening comprehension. Offloading the workload in the phonological loop by using the touchpad for input tasks does not result in an obvious benefit. If a non-navigation command is performed on the touchpad, the user needs to find the appropriate command menu on the touchpad, browse the menu to find the command, and click a button to execute the command. This is more effort than recalling the command and speaking it out. Therefore, speech input is chosen for non-navigation tasks.

13.1.2.2 Explanation by task match to modalities and touch-speech coordination. The touchpad input provides an advantage of mapping the information structure onto a tangible physical space and hence assists navigation. Navigation operations are therefore more frequently performed on the touchpad.

The speech input provides direct access to commands and saves time by avoiding menu browsing. Non-navigation commands are usually short and not performed continuously. As such, giving non-navigation commands using speech input is more efficient than searching for the commands on the touchpad. For the reasons of efficiency and simplicity, users naturally choose speech input for non-navigation commands.

Parallel tasks are performed using touch and speech separately to reduce the interference with each other. The subjects kept their fingers on the touchpad to retain their position in the information space and gave a short, one-time command using the speech input. They then resumed their previous navigation task quickly by continuing from the location that their finger retained on the touchpad. This touch-speech coordination helped to maintain a fluent command flow.

13.1.3. Resilience to Changing Modality Even under Error Pressure

The subjects did not change modality much, even when the error rate rose.

For sighted subjects, out of 1641 input operations that could be performed using a modality different from the one used in the previous input, only 222 (i.e., 13.52%) modality switches occurred. For visually impaired subjects, out of 5518 input operations that could possibly involve modality switching, only 755 (i.e., 13.68%) switches occurred. Hypothesis testing revealed that, in both the situation with low error rates and the situation with high error rates, when errors occurred, in significantly more occasions, the subjects continued to use the failing modality, rather than switching to another modality, for error correction.

The simple explanation is that changing the modality requires resetting the whole cognitive frame, and hence, requires more cognitive resources, which is avoided by users.

A longer explanation can be formed by looking at Broadbent's Bottleneck attention models (1958). The bottleneck theory specifies that the amount of information that can be processed and attended to by a human at any given time is limited. Therefore, concurrently performing tasks that compete for cognitive resources generally results in a

drop in performance for one or all tasks. When errors occur, users need to perform the following concurrent tasks: recognizing and understanding the input failure, finding solutions through more user inputs, and comprehending computer speech output. If switching modalities is to be performed, competition for cognitive resources among concurrent tasks will be more intense. Through practice the subjects already were aware that concurrently processing tasks could result in a drop in task performance. Therefore users will avoid modality switching for error correction unless switching has become a routine cognitive task that demands fewer resources.

13.1.4 Individual Differences

Although common multimodal usage patterns were supported by significance testing, individual differences existed. When error rates were not increased on purpose, five out of 19 visually impaired subjects used unimodal input. More specifically, one of them used speech input only, while the other four used touch input only. When error rates were increased by the experimenter, four of the five subjects switched input modalities, but one subject still insisted on touch input only.

There is not a simple answer as to why these users chose to use a single modality. By looking into the input operations they performed, the ratings they gave for the modalities in a follow up questionnaire, and their comments during the interview that followed their experiment session, the experimenter collected the following information, which provided more insight into their choices.

To illustrate these details, the subject IDs are used. The four subjects who insisted on the touch input during the session with low error rates are labeled S4, S14, S16 and

S17. The subject who used speech input only during the same session is labeled S8. S14 is the subject who insisted on touch only even when error rates were increased.

During the task sessions with low error rates, S4, S14, and S17 performed more navigation operations than the other subjects. The navigation operations performed by S8 were close to the average of all subjects. In accordance, S4, S14, and S17 performed fewer non-navigation operations than all other subjects, while the amount of non-navigation operations performed by S8 was close to the average. Performing more navigation operations than other subjects might have encouraged S4, S14, and S17 to use more touchpad input than other subjects.

By using unimodal input, all of the five subjects achieved lower error rates than the other subjects. S14, the subject who insisted on unimodal input even when the error rate was increased, achieved the lowest error rate among all subjects during the session with low error rates. Under a lower error pressure than others, the five subjects were not as motivated to switch modalities for error correction as others.

When error rates were increased, the five subjects no longer had the lowest error rates. Four of them started to switch modalities. Two of the four started to use multimodal error correction. But the other two of the four insisted on unimodal error correction instead of switching modalities to correct errors, despite having started to switch modalities for other tasks.

The comments given by the subjects during their interviews revealed why they had chosen one of the input modalities and stayed in it without switching. Each interview was conducted after the session with low error rates, but before the session with high error rates, so the subjects' answers were not influenced by the increased error rates.

S8, who used speech input only during the low-error rate session explained that "[Using the touchpad] is not as easy as saying it. You have to press more buttons and do more steps. When you verbally speak it, it will take you there in just one step." S17's comments represented the opinions of the four subjects who stayed in touch mode during the low-error rate sessions: "[speech and touch] both have a proper place for use, if you said which one ... like I have to have one, then I probably will take the touchpad. Even though it might be more frustrating to find where it is, for me I think once you learn how to use it, you could use it. ... because for some reason to think [what the speech command is for] next sentence [is], compared to just do the next sentence [on the touchpad], is an extra brain step, which takes a little longer to me. But if you have no use of hands, speech is excellent to read a newspaper."

However, these subjects' subjective ratings were not entirely consistent with their choice of modality. S4 did not fill out a questionnaire, so the discussion is based on ratings by the other four subjects.

For navigation operations, S8, S14 and S16's ratings on ease of learning, ease of use and likelihood to use were consistent with their usage of the modalities during the low error rate sessions – they rated the modality they each insisted on easier to learn, easier to use and that they were more likely to use it than the other modality that they did not choose. On the other hand, S17, who used touch only in the low error rate condition, rated speech easier to learn and use than touch, but he admitted that he would be more likely to use touch for navigation tasks.

For non-navigation operations, all of the four subjects rated speech easier to use than touch, despite the fact that three of them only used the touch input.

The above results imply, to some extent, that the distribution of the input operations used, in combination with error rates encountered, determined their individual preferences for one input modality over another. But the fact that their ratings were not consistent with their modality use implies that their choice of input modality was not entirely conscious. They naturally made a choice based on learned experience without much conscious thought.

13.1.5 Other Findings

The above findings were deemed the most research significant, in particular, because the results violate common practice or belief, e.g., that multimodal systems are good for error correction when input in one modality is failing. The findings listed below are also of interest but deemed less important.

13.1.5.1 Less Modality Switching in Tasks Demanding Higher Cognitive Resources. The type of cognitive task has been found to have a significant impact on users' multimodal input usage. When performing routine cognitive tasks, the subjects switched input modalities significantly more frequently than when they performed problem solving tasks.

Routine cognitive tasks are familiar tasks and processed automatically, while problem solving tasks require a higher level of cognitive resources. Research has pointed out that human attentional resources have limited capacity (Kahneman, 1973), and that bottlenecks exist along information processing stages (Broadbent, 1958). Humans divide attentional resources between time-sharing tasks to deal with time and resource competition between tasks. One rule for attention allocation is that when an automatic

processing task is combined with any other more cognitively demanding task, more cognitive resources are available for the later (Allport et al. 1972; Shaffer, 1975; Shiffrin, 1977; Sweller et al. 1990).

When routine cognitive tasks are processed, since they can be processed automatically, more attentional resources are available for other concurrent tasks, such as modality switching. When problem solving tasks are processed, since they demand a high level of attentional resources, resources available for modality switching become limited.

13.1.5.2 Increased Modality Switching for Error Correction due to Increased Error Rate. It is not surprising that error rates have significant effects on users' error correction strategies. When error rates are increased significantly, users switch their input modality more frequently to correct errors. A simple explanation for this is the following: When increased errors are encountered in the input modality that a user originally has chosen, the user eventually gives up on the modality of choice and switches to the other modality to avoid the errors.

However, users, in general, still prefer to stay in the same modality when errors occur, even though the frequency of their modality switches for error correction increases.

13.1.5.3 Increased Modality Switching, in General, due to Increased Error Rate. This research found that error rates significantly affected users' modality switching, in general, and not merely when error correction was needed. In short, increased error rates resulted in increased modality switches anywhere in the user task.

The reason for this could be the following: When error rates are increased, users' practice more modality switching for error correction. This makes modality switching more familiar, eventually becoming a routine cognitive task. Modality switching can then be performed with a higher level of automation, requiring less cognitive resources. Therefore users are able to switch modalities more often, in general, with less resource competition with other tasks.

13.1.5.4 Similar Multimodal Choices Exhibited By Sighted and Visually Impaired Users. Similar multimodal interaction patterns were found in the sighted and the visually impaired subjects' choices of input modalities based on input operation types, and the subjects' preferred strategy for error correction.

For both sighted and visually impaired subjects, navigation operations were usually performed using touchpad input, while non-navigation operations were usually performed using speech input. For both sighted and visually impaired subjects, correcting errors using the modality that was failing was significantly more prevalent than correcting errors by switching the modality.

13.1.5.5 Higher Usage of Modality Learned First. The exploratory study with sighted users found that when different input modalities of a multimodal system were taught separately, increased usage of the modality taught first could be observed. Among the subjects who were trained on the speech input first, there was no significant difference between the amount of speech and the amount of touch input used. Among the subjects who were trained on the touchpad input first, the touchpad input was used significantly more often than the speech input. It was also found that the subjects trained on the speech input first used significantly more speech input and less touch input than the subjects who

were trained on the touch input first. This indicates that training order has an effect on modality choice, but this primacy effect may disappear with use given the other reasons observed for modality choice.

13.2 Implications for the Design of an Eyes-Free Information Browser

Based on the findings in this research, suggestions can be made to guide the design of eyes-free systems accessing hierarchical text information sources.

Implement multimodal input. Most users use multimodal rather than unimodal input because (1) different input modalities provide different advantages, and (2) humans' attention and working memory process information distributed in different modalities more efficiently than processing all the information in one modality. Although the study was only performed using an information browser, the theory behind the results suggests that multimodal systems, in general, for a wide range of user interfaces would be a better design than unimodal input.

Implement modalities based on tasks. Implementing full functions in each modality is probably not necessary, because users will use a modality only for certain types of tasks. And users will stay in that initially chosen modality even if errors occur. Touch input on a tangible surface is appropriate for navigation operations. Speech input that allows direct access to functions is appropriate for non-navigation operations. By choosing the right modality for the right task, implementation efforts can be saved, computing resources can be used for other applications, and interface operation learning by the user reduced.

Allow modality switching for broken task flow. Intensive usability testing is needed during system design to determine whether there are task flows in which one type of operation is frequently interfered with by another type of operation. For example, whether browsing a command menu (a navigation task) is frequently interrupted by the task of increasing touchpad sensitivity (a non-navigation task). If tasks of different types constantly interfere with each other, modality switching should be allowed in these task flows.

Implement alternate methods in single modality to perform one task. The pilot study found that when the only way to fix an error was either repeating the failed command or switching to the other modality, the subjects repeated the failed command, but had low success rates. When there was more than one method in the failed modality for a problem fix, the subjects used the alternative methods in the failing modality rather than switching modalities. Thus, implementing alternate methods in a single modality for task performance matches a users' natural behavior.

Allow modality switching for critical error correction and train users to use it. If efficient error handling is critical during the use of a system, such as with an emergency management system, to avoid the problem of one modality failing completely making the system unusable, alternative input modalities should be implemented to allow error correction in other modalities. However, users have to be trained to use multimodal error correction because this research indicates that they will not readily switch modalities for the correction.

Train users first on the modality that is most appropriate for a given task. Since the modality taught first will be more frequently used in the future during users'

interaction with the multimodal system, when users are trained on performing certain type of tasks, they should be trained on the modality most suitable for the task type. Based on the results from this research, if both modalities are made available to users for certain design considerations, touch input should be trained on first for navigation tasks, while speech input should be trained on first for non-navigation tasks.

Do not create parallel speech and touch commands. Speech grammar should be different from touchpad grammar. Speech commands should be designed to best support direct access to functions, while touch commands should be designed to make navigation easy and efficient. When speech grammar is designed for navigation, it loses its advantage and will rarely be used. During this research, some speech commands were never used by the subjects, such as "settings menu", which points to the first setting on a list of audio settings, and "next/previous setting", which allows browsing the setting list. These commands mirrored the touchpad grammar, but were never used because the touchpad provides faster browsing and requires less working memory resources.

Choose appropriate touchpad sensitivity to filter out accidental touches. The touch input is more error-prone for visually impaired users than for sighted users, because visually impaired users explore the touchpad space by fingering it and can easily touch a spot on the sensing area that they did not mean to touch. This disrupts the current task. Since accidental touches are usually light, usability research needs to be done during system design to select a sensitivity of the touchpad that can filter out accidental touches. Alternatively, touchpad sensitivity can be a user setting for those who are heavy- or light-handed.

Implement flexible but not open speech grammar. One of the reasons why the subjects felt that speech input was less easy to learn and use was the rigidness of the speech grammar. For each speech input operation, a number of similar speech commands should be designed. Intensive usability research should be done during design to make the grammar close to naturally spoken commands. Open speech grammar refers to a grammar with which users speak naturally without using a fixed set of words and phrases. Such a grammar, however, has significantly lower recognition rates leading to other usability problems.

13.3 Contributions and Limitations

This research is the first one that investigated sighted and visually impaired users' non-visual multimodal interaction behavior in parallel. It has revealed important non-visual multimodal interaction patterns, explained the patterns using cognitive psychology theories in human attention and working memory, and discussed implications for the design of eyes-free information browsers.

In addition to discovering multimodal usage patterns, the research has also contributed in the following aspects:

- A speech grammar was designed for a non-visual information browsing system.
- A non-visual multimodal system was designed for hierarchical text browsing by both sighted and visually impaired users.
- A Wizard of Oz feature was created for simulating speech recognition functions and administering errors in speech and touch pad interaction for visually impaired users.
- Coding methods were developed to capture human modality usage from the experiment videotapes.

However, the research has its limitations.

The first limitation is the small sample size. A larger sample size may improve the significance of the results uncovered and also add power to conclusively reject those hypotheses not supported in this research.

The second limitation is that most subjects categorized as "with working vision" in this research might not have sufficient vision to be differentiated from the subjects categorized as "with no vision", because most subjects "with working vision" belonged to the legally blind category.

The third limitation is the interface used. It is not known how much of the results are from the specific interface design, that is, some tasks could have been inherently more difficult in one modality than another because of the design of the interface. For example, if keys on the interface had been available for the most commonly used speech commands, perhaps a larger number of users would have stayed with only the touch interface.

These limitations can be addressed in future research.

13.4 Future Research Directions

A range of research can be done in the future to expand understandings obtained from this research and to apply the results of this research to interaction design.

In the future, more participants based on recruiting standards that sampled a population more effectively could be invited to participate in the same controlled experiment. The addition of their data to the current data set may reveal stronger

statistical power which is expected to make the results more persuasive and also generalizable to a larger user population.

Users with low vision but not falling into the legally blind category should be recruited to complete the multimodal interaction comparison between users with usable vision and users without usable vision.

The research can be expanded to applications accessing various types of information structures and formats, not limited to the hierarchical textual information used in this research. Different information structures that could be included include hypertext and geospatial information, i.e., auditory street maps. Understanding how multimodal interaction could increase the accessibility of these information structures is of special value to the visually impaired users because of their more and more common use of the Internet and their need for non-visual guides for navigating their world.

A different information format that can be incorporated into the research is multimedia, such as computer games and digital music. These are becoming an important part of daily entertainment for visually impaired computer users. New research can focus on how multimodal interaction can make computer games more fun, and how multimodal interaction can make multimedia files more accessible.

The research can also be expanded to different computing platforms, not restricting to desktops and laptops, but including mobile devices such as PDA's and cell phones. Mobile devices have not been adopted by visually impaired users as broadly as they are by sighted users. This is mostly because of the accessibility issues residing in the design of mobile devices. Providing multimodal interaction on these devices is a potential way to improve a devices' accessibility. Mobile devices have a high potential to be

designed for increasing visually impaired users' mobility, which is the most desired capability by a high volume of visually impaired users. Mobile devices, hence, have a high importance in the future research agenda.

APPENDIX A

IRB APPROVAL FOR THE EXPLORATORY STUDY

The IRB approval for conducting the exploratory study is provided below.

NJIT NEW JERSEY INSTITUTE OF TECHNOLOGY

Institutional Review Board: HHS FWA 00003246
Notice of Approval
IRB Protocol Number: E12-04

Principal Investigators: Dr. Marilyn Tremaine

Title: Investigation of Speech and Touch Computer DIalogues

Performance Site(s): NJIT/Off Campus Sponsor Protocol Number (if applicable):

Type of Review: FULL [ ] EXPEDITED [X]

Type of Approval: NEW [ ] RENEWAL [X] MINOR REVISION [ ]

Approval Date: June 12, 2005 Expiration Date: June 11, 2006

1. **ADVERSE EVENTS:** Any adverse event(s) or unexpected event(s) that occur in conjunction with this study must be reported to the IRB Office immediately (973) 642-7616.

2. **RENEWAL:** Approval is valid until the expiration date on the protocol. You are required to apply to the IRB for a renewal prior to your expiration date for as long as the study is active. Renewal forms will be sent to you; but it is your responsibility to ensure that you receive and submit the renewal in a timely manner.

3. **CONSENT FORM:** All subjects must receive a copy of the consent form as submitted. Copies of the signed consent forms must be kept on file with the principal investigator.

4. **SUBJECTS:** Number of subjects approved: 65.

5. The investigator(s) did not participate in the review, discussion, or vote of this protocol.

6. **APPROVAL IS GRANTED ON THE CONDITION THAT ANY DEVIATION FROM THE PROTOCOL WILL BE SUBMITTED, IN WRITING, TO THE IRB FOR SEPARATE REVIEW AND APPROVAL.**

Dawn Hall Apgar

Dawn Hall Apgar, PhD, LSW, ACSW, Chair IRB June 12, 2005

APPENDIX B

IRB APPROVAL FOR THE CONTROLLED EXPERIMENT

The IRB approval for conducting the controlled experiment is provided below.

NJIT NEW JERSEY INSTITUTE OF TECHNOLOGY

Institutional Review Board: HHS FWA 00003246
Notice of Approval
IRB Protocol Number: E12-04

Principal Investigators: Marilyn Tremaine, Information Systems

Title: Investigation of a Speech and Touch Computer Interface

Performance Site(s): NJIT/Off-Site Sponsor Protocol Number (if applicable):

Type of Review: FULL [ ] EXPEDITED [X]

Type of Approval: NEW [ ] RENEWAL [X] MAJOR REVISION [ ]

Approval Date: May 8, 2006 Expiration Date: May 7, 2007

1. **ADVERSE EVENTS:** Any adverse event(s) or unexpected event(s) that occur in conjunction with this study must be reported to the IRB Office immediately (973) 642-7616.

2. **RENEWAL:** Approval is valid until the expiration date on the protocol. You are required to apply to the IRB for a renewal prior to your expiration date for as long as the study is active. Renewal forms will be sent to you; but it is your responsibility to ensure that you receive and submit the renewal in a timely manner.

3. **CONSENT:** All subjects must receive a copy of the consent form as submitted. Copies of the signed consent forms must be kept on file with the principal investigator.

4. **SUBJECTS:** Number of subjects approved: 65.

5. The investigator(s) did not participate in the review, discussion, or vote of this protocol.

6. **APPROVAL IS GRANTED ON THE CONDITION THAT ANY DEVIATION FROM THE PROTOCOL WILL BE SUBMITTED, IN WRITING, TO THE IRB FOR SEPARATE REVIEW AND APPROVAL.**

Dawn Hall Apgar

Dawn Hall Apgar, PhD, LSW, ACSW, Chair IRB May 8, 2006

APPENDIX C

CONTROLLED EXPERIMENT – CONSENT FORM

The consent form used in the controlled experiment is provided below.

NEW JERSEY INSTITUTE OF TECHNOLOGY
323 MARTIN LUTHER KING BLVD.
NEWARK, NJ 07102

CONSENT TO PARTICIPATE IN A RESEARCH STUDY

TITLE OF STUDY:

Investigation of a Speech and Touch Computer Interface

RESEARCH STUDY:

I,_____, have been asked to participate in a research study under the direction of Dr. Marilyn Tremaine. Other professional persons who work with them as study staff may assist to act for them.

PURPOSE:

The purpose of this study is to investigate the use of speech and tactile computer dialogues.

DURATION:

My participation in this study will last for at total of 4 to 6 hours to be completed over 2 consecutive days.

PROCEDURES:

I have been told that, during the course of this study, the following will occur:

On the first day:

- The experimenter will read the study introduction.
- The subject will fill out this consent form.
- The subject will fill out a background questionnaire which collects the information of the subject related to the study.
- The subject will participate in a tutorial to learn to use the speech and the touch input methods. The tutorial will be followed by a practice session.

On the second day:

- The subject will participate in a warm-up session to practice speech and touch input operations learned on the first day.

- The subject will participate in an experiment session during which s/he will finish a series of tasks using the provided input methods.
- The subject will be interviewed about their use experience with the system and fill out a post-questionnaire.

PARTICIPANTS:

I will be one of about 65 participants to participate in this trial.

RISKS/DISCOMFORTS:

There may be risks and discomforts that are not yet known.
I fully recognize that there are risks that I may be exposed to by volunteering in this study which are inherent in participating in any study; I understand that I am not covered by NJIT's insurance policy for any injury or loss I might sustain in the course of participating in the study.

CONFIDENTIALITY:

I understand confidential is not the same as anonymous. Confidential means that my name will not be disclosed if there exists a documented linkage between my identity and my responses as recorded in the research records. Every effort will be made to maintain the confidentiality of my study records. If the findings from the study are published, I will not be identified by name. My identity will remain confidential unless disclosure is required by law.

VIDEOTAPING/AUDIOTAPNG:

I understand that I will be video and audio taped during the course of this study. Video and audio tapes will be stored for 3 years after the end of this project which is June 2006. Three years after the end of the project the tapes will be erased. The tapes will be stored in a locked office at NJIT and will not be made available to anyone except the investigators including Dr. Marilyn Tremaine, Xiaoyu Chen, Robert Lutz, and John Visicaro who are involved in this research.

PAYMENT FOR PARTICIPATION:

I have been told that I will receive no compensation for my participation in this study.

RIGHT TO REFUSE OR WITHDRAW:

I understand that my participation is voluntary and I may refuse to participate, or may discontinue my participation at any time with no adverse consequence. I also understand that the investigator has the right to withdraw me from the study at any time.

INDIVIDUAL TO CONTACT:

If I have any questions about my treatment or research procedures, I understand that I should contact the principal investigator at:

Dr. Marilyn Tremaine
Information Systems Department
New Jersey Institute of Technology

Newark, New Jersey 07102
Email: tremaine@njit.edu
Telephone: 973-596-5284

If I have any addition questions about my rights as a research subject, I may contact:

Dawn Hall Apgar, PhD, IRB Chair
New Jersey Institute of Technology
323 Martin Luther King Boulevard
Newark, NJ 07102
(973) 642-7616
dawn.apgar@njit.edu

SIGNATURE OF PARTICIPANT

I have read this entire form, or it has been read to me, and I understand it completely. All of my questions regarding this form or this study have been answered to my complete satisfaction. I agree to participate in this research study.

Subject Name: _____

Signature: _____

Date: _____

SIGNATURE OF READER FOR PARTICIPANTS WHO ARE VISUALLY IMPAIRED

The person who has signed above, _____, has visual impairment, I read English well and have read for the subject the entire content of this form. To the best of my knowledge, the participant understands the content of this form and has had an opportunity to ask questions regarding the consent form and the study, and these questions have been answered to the complete satisfaction of the participant (his/her parent/legal guardian).

Reader Name: _____

Signature: _____

Date: _____

SIGNATURE OF INVESTIGATOR OR RESPONSIBLE INDIVIDUAL

To the best of my knowledge, the participant, _____, has understood the entire content of the above consent form, and comprehends the study. The participants and those of his/her parent/legal guardian have been accurately answered to his/her/their complete satisfaction.

Investigator's Name: _____

Signature: _____

Date: _____

APPENDIX D

CONTROLLED EXPERIMENT – STUDY INTRODUCTION

The following introduction of the research to the participants was used at the beginning of the controlled experiment.

Study Introduction

Dear subject,

Thank you very much for agreeing to participate in the study on AudioBrowser. AudioBrowser is a system that will read news for you during the study. You are asked to use a touchpad and your speech to operate AudioBrowser. Your preferred ways to operate the system and your evaluation will help up design better speech and touch input mechanism for non-visual information systems.

The study will be conducted in two separate sessions on two consecutive days. In the first day you will participate in a tutorial that teaches you how to use AudioBrowser. You will get hands-on experience on using the system. On the second day you will participate in an evaluation session, during which you will be asked to perform a list of tasks using AudioBrowser. We will catch problems in the system design and your preferred ways of operation during this session. This session will be videotaped for later analysis. Following the evaluation session, we will interview you about your experience with AudioBrowser. We will also collect some background information about you that helps us create a profile of the user group. Your participation in each day will be approximately 2 hours.

Please feel free to ask questions or provide opinions at any time during the study. Please understand that this is an initial design of the system and it could have problems. If you have difficulties using the system, it is due to design problems that we need to catch and fix. We appreciate the time and efforts you devote to help with designing better non-visual interfaces.

We are now giving you the detailed study procedure and a consent form to sign. Note that you are free to quit this study at any time.

APPENDIX E

CONTROLLED EXPERIMENT – BACKGROUND QUESTIONNAIRE

The following background questionnaire was used to collect the participants' information related to the research.

Background Questionnaire

The following questions ask you to tell us something about your background. This information will help us to understand the system design needs for different user groups. All information is completely confidential. You are free to decide not to answer any specific question.

Name:_____ Today's Date: _____

_____

For the following set of questions, please indicate which of the answers best applies to you.

1. Which age group do you belong to?

   a. ( ) 20 – 29 b. ( ) 30 – 39 c. ( ) 40 – 50

2. Could you tell me which education level applies to you?
   a. ( ) high school b. ( ) 1-2 years university
   c. ( ) 3-4 years university d. ( ) advanced education beyond bachelor's degree

3. At what age did you become visually impaired? _____

4. What is your current vision? _____

5. For how many years have you been visually impaired? _____

6. What caused your vision impairment? _____
   We ask this question because some of the factors associated will affect our design decision. For example, people whose vision problem was caused by diabetes may be affected in their sense of touch, and hence we should carefully design the touch input for them.

7. For how many years have you been using a computer? _____

8. Describe the current setup you use to access information on your computer:

   What software do you use to access information and computer programs? For example, JAWS.

_____

_

How do you give commands to your computer?

a. Regular keyboard
b. Braille input
c. Other input (Please explain)

_____

How does your computer give you its output?

a. Speech output
b. Braille output
c. Other output (Please explain)

_____

1. On average, how many hours do you use a computer every day? _____.

2. Now you will hear a list of tasks you might do with a computer. Please say "yes" if you use a computer to do the task.

_____ Create text documents
_____ Manage text documents
_____ Read news
_____ Write emails
_____ Use an online-chat program
_____ Post messages on online bulletin board
_____ Manage personal information such as contacts and appointments
_____ Search the Internet
_____ Develop web pages
_____ Write software programs

10. Do you use a computer for any other activities besides the ones I have indicated? If so, what are they?

_____
_____
_____

11. The following is a list of methods by which people access the news. For each method, please indicate whether you use it (1) often (2) somewhat (3) never.

a. Radio	(1) often (2) somewhat	(3) never
b. Television	(1) often (2) somewhat	(3) never
c. Newspaper web sites	(1) often (2) somewhat	(3) never
d. Braille newspapers	(1) often (2) somewhat	(3) never
e. Other (please explain)		
_____	(1) often (2) somewhat	(3) never
_____	(1) often (2) somewhat	(3) never
_____	(1) often (2) somewhat	(3) never

12. Please provide some web sites you visited on the Internet:

APPENDIX F

CONTROLLED EXPERIMENT – INPUT MODALITY TUTORIAL

The following is the tutorial teaching the participants how to use the speech and touch input modalities.

AudioBrowser Tutorial

Now you are participating in a tutorial session that teaches you how to use AudioBrowser. It starts with an introduction to the AudioBrowser system.

AudioBrowser is a program that lets you browse and read information through touch and speech. We are going to work mostly with newspapers. Audiobrowser will bring up the news sections of a newspaper first, such as international section, national section, sports section, etc. It will then let you select a news section based on your interests. Once you select a news section, the news articles will be available for you to read.

You will communicate with Audiobrowser via finger touches and voice. That is, you will operate the system using a touchpad and speech commands. AudioBrowser will talk back to you.

Through this tutorial you will learn to use the touchpad and the speech commands that AudioBrowser understands. Note that AudioBrowser sometimes makes mistakes. The sensitivity of the microphone, sounds from the environment, and echoes in the experiment room are all possible causes of system recognition mistakes. In case a mistake occurs, you may either repeat your command or use other available commands to recover from the mistake.

This tutorial includes the following sections: (1) an overview of the touchpad input and the speech input, (2) a description of the functions of AudioBrowser and the ways to use those functions using the touchpad and the speech input. Whenever you have questions during the tutorial, please feel free to ask. You are also encouraged to try out the system when the experimenter is explaining a function.

An overview of the touchpad and the speech input:

In front of you is a touchpad used for controlling AudioBrowser. (Let subject feel the touchpad.) It is a rectangular device with an indented area at the center. The indented area can detect your touch and so we call it the sensing area. The sensing area is divided into three tracks (Guide the subject to feel the tracks). The news sections and articles and the system commands are mapped onto the tracks. When a news section or a command is

touched on the touchpad, the system speaks to tell you what is being touched. Beside the tracks, on the left and the right sides of the touchpad are two buttons (Guide the subject to feel the buttons). You then can click one of the two buttons to enter the section or to execute the command.

The speech input commands execute functions that are also available on the touchpad. So, whenever you want the system to do a task you can choose between the touchpad input and the speech input. Place the microphone in front of you, stay close to the microphone, and speak to it to let the system hear you. When hearing a speech command the system repeats what it hears and executes the command.

AudioBrowser functions and how to use the functions

1. Browse news categories and articles

When the system is started, the news categories of an available newspaper will be ready for you to read. Using the touchpad, you can browse the categories by gliding your finger on the top track from left to right. Now please try it (Guide the subject's finger). An example of the system's speech output is "International section, 3 articles, 2 subsections, item 1 of 8." Here *international section* is the news category; "3 articles and 2 subsections" indicates that there are 3 articles and additional 2 subsections in the international section. "Item 1 of 8" indicates that there are a total number of 8 news categories and *international section* is the first one. Between two adjacent news categories you can hear a "click" sound that indicates the boundary between the two. Now please use the touchpad to browse the available news categories again.

You can also browse the news sections using speech input. These are the commands you will use: "***next category***," "***next article***," and "***next item***." These commands let you go to the next available news section or article. Although these commands are worded differently, they are equal to the system. Similarly, you can use the commands: "***previous category***," "***previous article***," and "***previous item***" to go to the previous available news section or article. These commands, again, mean the same. Now please try these commands.

> Task 1:
> Please find the Sports Section using speech commands. Then find the Business Section using the touchpad.

When you find a category interesting to you, you can enter the category to read the articles inside. To enter a category using the touchpad, locate the news category on the top track first, then click the right button. Now please try to enter the national section. When you clicked the right button, you heard the system read the title of the first article, the author of the article, and the article's order among the total number of articles in this category. The top track is now changed – Instead of having the general news categories such as "international section," "national section", it now has the articles and the subsections in the news section you just selected. You can listen to the titles of these

articles by gliding your finger on the top track. Now try this. To exit this category, click the left button. Please try it. Now listen to the items on the top track. It comes back to the general news categories.

You just learned to select a news category and exit a news category using the touchpad. By using speech input you can do the same thing. The speech commands to select a news category are "***select***" and "***zoom in***," which work equally well. The commands to exit a news category are "***exit***" and "***zoom out***," which work equally well, too. After you have entered a news category, you can use the commands "***next article***," "***next item***," or "***next category***," and "***previous article***," "***previous item***," or "***previous category***" to browse the news articles and sub news sections inside this category. Now please try these commands.

Task 2:

- Please use speech commands to go to the National Section and use the touchpad to zoom in.
- Please use the touchpad to zoom out from the National Section, find the Technology Section, and use a speech command to zoom into the Technology Section. Finally, use a speech command to zoom out.

2. Read an article

You have learned to browse news sections and the titles of news articles. I am now going to explain how to read an article when you hear an interesting title. Again you can use either the touchpad or the speech input.

When using the touchpad, locate the article you want to read using the top track first. Then you will move to the middle track. On the middle track of the touchpad, there is a list of text units. Now glide your finger on the middle track from left to right slowly. You just hear four commands: "set to word," "set to sentence," "set to paragraph," and "set to complete article." By touching "set to sentence" you request the system to read the article sentence by sentence. By touching "set to paragraph" you request the system to read the article paragraph by paragraph. The system will pause after each sentence or paragraph. The command "set to complete article" allows you to read the whole article without stop, unless you stop it. The "Set to word" command will be useful when you are searching for a word to spell.

Speech commands can do the same tasks. The speech commands to use are: "*set to word*," "*set to sentence*," "*set to paragraph*," and "*set to complete article*." If you forget these text units, you can use the command "*output unit*" to have the system tell you the text units. (Please try these speech commands.)

Now let me brief what you have done:
(1) First, you locate an article. To locate an article you glide your finger on the top track or use speech commands: "next article" "previous article" "next item" "previous item" "next category" "previous category;"

(2) Second, you select the text unit by which you want the system to read the article. On the touchpad you do so by gliding on the middle track. Using the speech input you use the commands "set to word, sentence, paragraph, or complete article."

Now what's the next step? Again you have two choices, the touchpad or the speech input. On the touchpad you can click either the right button to read the next text unit, or the left button to read the previous text unit. But pay attention that if you have set the text unit to the complete article, only press the left button, which will read the whole article from *the beginning*. The right button doesn't work. To pause reading at any time, press the left and the right buttons together. Now please try to let the system read by "sentence," then by "paragraph," and then by "complete article."

Note that the two buttons now have different functions – For the top track on which you browse news categories and article titles, the two buttons are to enter or exit a section. For the middle track on which you set the text unit, the two buttons are to read the next or previous text unit.

You clicked the touchpad buttons to read the next or previous text unit. You can use speech commands to do the same thing. The speech commands to use are: "*next word*," "*previous word*," "*next sentence*," "*previous sentence*," "*next paragraph*," "*previous paragraph*," "*read article*" and "*pause*." The command "*read article*" will read an article from *the beginning*. Now please try these commands.

A trick here is that the speech commands are more flexible than the touchpad – you don't have to say "set to sentence" before using the command "next sentence." You can use the command "next sentence" directly at any time. Similarly, you can use the commands "*next word*", "*previous word*", "*next paragraph*" and "previous paragraph" directly without setting the reading unit to word or paragraph first. Now please say "next word" first, and then say "next sentence."

To resume reading after a pause, say "*resume*", which will resume reading from where it was paused. An alternative way to resume reading is to use the touchpad. Go to the middle track first and select the text unit you want the system to read by, and click the right button to resume from the next text unit, or click the left button to resume from the previous text unit. Now please try these commands.

Task 3:

- Please use a speech command to go to the next article in the current news category.
- Use the touchpad to have the system read the next four sentences. While the fourth sentence is being read, use the touchpad to pause in the middle of the sentence. Then use the speech input to resume from where it is paused.
- Use a speech command to let the system read the next paragraph. In the middle of the paragraph, use a speech command to pause reading. Then use the touchpad to resume reading from the next paragraph.

When you need to spell a word, use the speech commands "***spell***" or "***spell word***." These commands spell the last word in the article read by the system. To spell using the touchpad, set the text unit to word first, then press the left and the right buttons together. Of course you can set to word using either the middle track of the touchpad or using the speech command "set to word." Now please try these commands.

When you need to repeat a text unit in the article last read by the system, use the speech command "***repeat***," or press the right button to go the next text unit and the left button to return to the unit you want to repeat.

3. Adjust audio settings

You have learned how to read an article. What if you need to decrease the reading speed or increase the reading volume?

AudioBrowser allows you to change 5 audio settings: reading speed, volume, the voice used to read the article, the pitch of the voice, and the volume of the non-speech audio feedback, e.g., the clicks heard between news items. These settings are on the bottom track on the touchpad. Listen to these settings by gliding your finger on the bottom track. At this time clicking the right button will increase the value of the setting you touched last, and clicking the left button will decrease the value of that setting. Now please try these controls.

To change the settings using speech input, use the following speech commands: "***increase speed***," "***decrease speed***," "***increase volume***," "***decrease volume***," "***next voice***," "***previous voice***," "***increase pitch***," "***decrease pitch***," "***increase tone volume***," and "***decrease tone volume***." Now please try these controls. **When changing the voice, please only use the first three voice, Mary, Mike, and Sam.**

If you forget the audio settings available, there are three ways to remind yourself. The first way is to glide your finger on the third track to browse the settings. The second way is to use a speech command "***settings menu***," which has the system tell you all the settings available for adjustment. The third way is to give the following speech commands in order: "***first setting***," "***next setting***," and "***next setting***" until the setting that you are looking for is reached. Now please try these controls.

After adjusting the audio settings, there are multiple ways to return to the article being read. Using speech input, you can say "resume," "next sentence / paragraph," "previous sentence / paragraph." You can also use the command "read article" to read from the beginning of the article.

To return to the article using the touchpad, go back to the middle track to select a text unit and click the right button to read the next text unit, or click the left button to read the previous text unit. Now please try these controls.

If you don't want to return to the point where you paused within the article, but to read other articles or news categories, use the speech commands "*next article*," "*next category*," etc., which you have used to browse news categories and articles. You can also place your finger back to the top track to locate the article titles or news sections. Now please try these controls.

Task 4:

- Please use a speech command to go to the next article in the current news category.
- Please use a speech command to have the system read the next paragraph. Pause in the middle of the paragraph using speech.
- Use the touchpad to decrease the reading speed by one level.
- Use the touchpad to resume reading from the next paragraph. Pause in the middle of the paragraph using the touchpad.
- Please use speech commands to increase the reading pitch by two levels.
- Use speech to resume reading from where it has paused.

You have learned all the functions and controls of AudioBrowser. Now we will have a short break (5 minutes). After the break I will briefly summarize the speech and touchpad controls you have learned, and give you a list of tasks to practice.

APPENDIX G

CONTROLLED EXPERIMENT – TASK SHEET FOR PRACTICE IN DAY ONE

The following tasks were used in the practice session held after the input modality tutorial in day one.

Practice of Speech and Touchpad Input

In this practice session you will finish a list of tasks using the speech and the touchpad input you have learned. When there is a question or a problem please let the experimenter know. The experimenter will help you. This practice session will be video taped. You are suggested to use about 30 minutes to finish all the tasks.

Task 1.
Please use speech commands to finish the following steps:
- Find and enter the Sports Section.
- Find an article titled "Johnson Wins 200 Meters Semifinal."
- Set the reading unit to sentence and read the next sentence.
- Read the next paragraph and pause in the middle of the paragraph.
- Change the voice to the next available voice and continue to read the article. Wait until the system finish reading the current text unit.
- Go back three words and spell the word.
- Please read two more sentences.
- Exit the Sports Section.

Task 2.
Please use the touchpad to perform the following steps:
- Find the Politics Section and enter the section.
- Find an article titled "Bush Is Seeking Safe and Solid Running Mate."
- Go to the third paragraph. Pause when the third paragraph is being read.
- Go back five words and spell the word.
- Change the voice to the previous voice and read the next sentence.
- Read one more paragraph and exit the Politics Section.

Task 3.
- Please use the speech input to go to and enter the International Section.
- Use the touchpad to find the subsection titled "Europe" and enter it. The "Europe" section can be a subsection.

- Use the touchpad to find an article titled "Spain Suspects Basque Group in 2 Attacks."
- Use the speech input to read the next four sentences.
- Use the touchpad to read the next five words.
- Use the touchpad and the speech input to spell the last word respectively.
- Use the touchpad to decrease the reading speed by one level and use the speech input to read the next sentence.
- Use the speech input to increase the reading volume by one level and use the touchpad to resume reading.
- Please repeat the last sentence using speech and touch respectively.

Task 4.

Perform the following steps. Please use any mixed speech and touch input that you like:

- Find a news article interesting to you. The article should be in a different news category.
- Spell the name of the author and read five more sentences.
- Increase the tone volume by 2 levels and read the next paragraph. Pause reading in the middle of the paragraph.

APPENDIX H

CONTROLLED EXPERIMENT – TASK SHEET FOR WARMING-UP IN DAY TWO

The following tasks were used as a warming-up practice before the controlled experiment in day two.

Warm-Up Tasks

In this practice session you will finish a list of tasks using the speech and the touchpad input you have learned yesterday. When there is a question or a problem please let the experimenter know. The experimenter will help you. This practice session will be video taped. You are suggested to use about 15 minutes to finish the tasks.

Task 1.
Please use speech input to finish the following steps:
- Please find and enter the *New York Times* Section.
- Find and enter the *National* Section.
- Find an article titled "*Supreme Court Hears Case on Abortion Rights.*"
- Read the next paragraph.
- Read the next paragraph and pause in the middle of the paragraph.
- Decrease the reading speed by one level and continue to read the article. Wait until the system finishes the current paragraph.
- Go back three words and spell the word.
- Read one more sentence.
- Exit the *National* Section.

Task 2.
Please use the touchpad to perform the following steps:
- Find the *Education* Section.
- Enter the *Education* Section and find an article titled "*Learning-Disabled Students Blossom in Blended Classes.*"
- Go to the third paragraph. Pause when the third paragraph is being read.
- Go back four words and spell the word.
- Increase the reading volume by one level and read the next sentence.
- Read one more paragraph and exit the *Education* Section.

Task 3.

Perform the following steps. Please use any mixed speech and touch input that you like:

- Find the *International* Section.
- Find the news about *Mexico's Leader's Attitude toward Migration.*
- Spell the last name of the author and read five more sentences. Pause in the middle of the fifth sentence.
- Change the voice to the next available voice and increase the reading speed by one level. Resume reading.
- Repeat the last sentence and read one more paragraph.

APPENDIX I

CONTROLLED EXPERIMENT – EVALUATION OF PARTICIPANTS' ABILITY TO UNDERSTAND COMPUER SYNTHESIZED SPEECH

The following articles are TOEFL exam samples from ETS Web site. They were used to evaluate the participants in the controlled experiment on their ability to understand computer synthesized speech output. The articles are marked using the scripts read by the AudioBrowser system.

(Articles are marked with scripts read by AudioBrowser)

<np title=|TOEFL Listening Sample Questions| date=|December 1, 2005|>
<s name=|Lectures About the Nature|>
<a hl=|Article One|
bl=|By ETS|
t=|Today's discussion is about a common animal reaction — the yawn. The dictionary defines a yawn as "an involuntary reaction to fatigue or boredom." That's certainly true for human yawns, but not necessarily for animal yawns. The same action can have quite different meanings in different species.
For example, some animals yawn to intimidate intruders on their territory. Fish and lizards are examples of this. Hippos use yawns when they want to settle a quarrel. Observers have seen two hippos yawn at each other for as long as two hours before they stop quarreling.
As for social animals like baboons or lions — they yawn to establish the pecking order within social groups, and lions often yawn to calm social tensions. Sometimes these animals yawn for a strictly physiological reason — that is, to increase oxygen levels. And curiously enough, when they yawn for a physical reason like that, they do what humans do — they try to stifle the yawn by looking away or by covering their mouths. |
>
<a hl=|Questions Based on Article One|
bl=|Each question is a paragraph. Please set the text unit to paragraph to read them. |
t=|Question 1: What is the speaker's main point?
Answer A:Animals yawn for a number of reasons.
Answer B:Yawning results only from fatigue or boredom.
Answer C:Human yawns are the same as those of other animals.
Answer D:Only social animals yawn.
Question 2: According to the speaker, when are hippos likely to yawn?
Answer A: When they are swimming.
Answer B: When they are quarreling.
Answer C: When they are socializing.

Answer D: When they are eating.
Question 3: What physiological reason for yawning is mentioned?
Answer A: To exercise the jaw muscles.
Answer B: To eliminate fatigue.
Answer C: To get greater strength for attacking.
Answer D: To gain more oxygen. |
>
<a hl=|Article Two|
bl=|By ETS|
t=|Now listen to part of a lecture from a gemology class.
In last week's lesson about the difference between metals and gems, we discussed how pliable true gold is. Today we are going to be talking about the diamond, the hardest known natural mineral. As most of you know from our introductory chapter, diamonds are the transparent form of pure carbon. Carbon crystals form deep in the Earth's mantle when high temperatures and extreme pressure occur. The term "diamond" comes from the Greek word adamas, which means unconquerable. In the jewelry business, diamonds are valued according to a few categories, known as the 4 C's. The cost of a diamond depends on its carat, color, cut, and clarity. Besides Africa, there are few areas around the world with large diamond deposits. However, diamond replication is a new trend that threatens the multimillion dollar industry. Researchers have discovered a way to produce large volumes of diamonds by putting carbon under extreme heat and pressure. This process causes the carbon to crystallize into diamonds. Even the trained eye cannot detect the difference between a natural diamond and one that is manufactured in this way. While this innovation could devastate the jewelry industry, it could also turn the precious stone into a common semiconductor. Not only are diamonds incredible conductors of heat, they are also efficient electrical insulators. Tremendous heat can pass through a diamond without causing any significant damage. |
>
<a hl=|Questions Based on Article Two|
bl=|Each question is a paragraph. Please set the text unit to paragraph to read them.|
t=|Question 1: What is the purpose of this lecture?
Answer A: To compare diamonds and gold.
Answer B: To discuss types of gems.
Answer C: To discuss the formation of diamonds.
Answer D: To review the elements of carbon.
Question 2: Which of the following is not one of the 4 C's used by the jewelry business?
Answer A: Carbon.
Answer B: Carat.
Answer C: Color.
Answer D: Cut.
Question 3: Where do natural diamonds form?
Answer A: In a manufacturing plant.
Answer B: In an electrical insulator.
Answer C: Deep in the Earth's mantle.
Answer D: Alongside metals such as gold.
Question 4: According to the professor, what are diamonds good for besides jewelry?

Answer A: They can create heat.
Answer B: They can hold heat in.
Answer C: They can damage insulators.
Answer D: They can conduct electricity. |
>
</s>

APPENDIX J

CONTROLLED EXPERIMENT – EXPERIMENT TASK SHEET

The document below is the task sheet used for the experiment sessions.

Experiment Task Sheet

You are being asked to do the following tasks. You can use the speech input and the touch input freely. A video camera will be used to catch your use of the system and any problems occurring.

You will do these tasks independently without help from the experimenter. All tasks are to be completed. If you encounter a problem, please tell the experimenter what the problem is.

The experimenter will read each task for you. If you do not hear a task clearly, feel free to ask the experimenter to repeat. Tell the experimenter when you finish a task so that the experimenter will then read the next task for you.

Task 1. (Finding the user's comfortable reading speed)

1) Please listen to all of the sections available at the top level.
2) Go to *National section* in *New York Times* section and find the article titled "*Busiest Hurricane Season on Record Ends*".
3) Set the reading unit to *Paragraph*, and have the system start to read the article.
4) Adjust the reading speed to a level comfortable for you. Do this by adjusting the reading speed, listening to the article and readjusting the reading speed until you feel it is comfortable. When you find the best speed for you, stop reading.

Task 2. (Testing the user's ability to understanding synthesized computer speech output)

Now you will listen to two articles. For each article, once the system starts to read it, it will not stop until the whole article is finished. Please try your best to understand the article. When each article is finished, you will answer some questions based on the article. When you are ready, the experimenter will have the system start reading.

(The experimenter has the system read articles in "Lectures About the Nature" section.)

The article is finished. Now you are to answer some multiple-choice questions based on the article. Listen to each question and the four choices of answers following it. They are labeled Answer A, Answer B, Answer C, and Answer D. You can ask the experimenter

to repeat any question and answer choice if you need. When you decide your answer to a question, tell the experimenter your answer and proceed to the next question. Now you can start listening to the questions.

Article 1			Article 2	
Participant's answer	Standard answer		Participant's answer	Standard answer
1)	A		1)	C
2)	B		2)	A
3)	D		3)	C
			4)	B

Task 3. (Routine cognitive tasks)

1) Please find the *Times Magazine Special Issue* Section which is at the top level.
2) Set the reading unit to *Sentence*. Enter the *Times Special Issue* Section.
3) Read all of the item titles available in this section.
4) Go to *Part One* of the article and start reading it. When four sentences are finished, pause in the middle of the fifth sentence.
5) Increase reading volume and resume reading. Then pause in the middle of the next sentence.
6) Have the system spell out the last word that the system read.
7) Go back four words and spell the word.
8) Resume reading, increase the reading pitch by three levels. And resume reading.
9) Decrease the reading pitch by two levels. Then start reading the article from the beginning. Pause after you have heard the words, "The TIME 100."

Task 4. (Comprehending a short article)

Now read the article again, from the beginning to the end. When you finish the article, finish three questions asked based on the article. The questions are in the section next to the article, titled *Questions based on Part One*. You will have the system read the questions for you and go back to the article to find the answers.

You can read the questions first if you want. You can repeat reading any part of the article or the questions as needed.

Task 5. (Comprehending a longer article)

Now you will read Part Two of the article. Answer the questions asked based on Part Two. Again you will have the system read the questions for you and go back to the article to find the answers.

Again you can read the questions first if you want. You can repeat reading any part of the article or the questions as needed.

Task 6. (Observing the user's error-handling strategies)

In this task section you will handle errors. The system will generate errors for you to handle. When an error occurs you can use any way to fix it. All errors are to be fixed.

1) Please tell the system to go to the New York Times Section, and enter the Health Section.
2) Please go to the third article
3) Please set the reading unit to "sentence" and read the next three sentences. Then repeat the last sentence.
4) Please command the system to read by complete article, and resume reading from where it was paused. After it reads two or three sentences, pause reading.
5) Please spell the last word read by the system. Go back two words and spell.
6) Please set the reading unit to "paragraph" and resume reading from where it was stopped. Wait until the system stops.
7) Please go to the next article.
8) Please increase the reading pitch by two levels and resume reading.
9) Please command the system to exit *Health Section* and zoom into *Business Section* that's two sections before *Education Section*.
10) Please command the system to decrease the reading speed by one level.

APPENDIX K

CONTROLLED EXPERIMENT – POST QUESTIONNAIRE

The following is the questionnaire used between the experiment session with low error rates and the experiment session with high error rates.

Post Questionnaire

Before you perform the last task, we would like you to tell us your opinions about the speech input and the touchpad input. You will give us your opinions for each of the tasks that we describe.

Please describe how would you do the following tasks using speech and using touch respectively?

1. Browse news sections and article titles
Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

2. Enter a news section.
Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

3. Exit a section

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

4. Set reading unit

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

5. When the reading unit is set to "paragraph", read the next paragraph

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

6. When the reading unit is set to "word", read the next paragraph

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

7. Read the next five sentences continuously

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

8. Pause reading

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

9. Resume reading

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

10. When the system reading has been paused, spell the last word that the system has read

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

11. When the system reading has been paused, find a word in the middle of a sentence

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

13. When the system is reading, decrease the reading speed

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

14. After the reading speed has been adjusted, resume reading

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

15. Repeat the sentence that was just read by the system

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

16. Get to know what reading units are available

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

17. Get to know what audio settings are available

Please describe how would you do this task using speech?

How likely would you use speech to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use speech to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

Please describe how would you do this task using touch?

How likely would you use touch to finish this task?
Likely 1 – 2 – 3 – 4 – 5 Unlikely

How easy would you find it is to use touch to finish this task?
Easy 1 – 2 – 3 – 4 – 5 Difficult

QUESTIONS FOR USER INTERFACE SATISFACTION [2]:

Overall reaction to the Speech Input:

Terrible	1 – 2 – 3 – 4 – 5	Wonderful
Difficult	1 – 2 – 3 – 4 – 5	Easy
Frustrating	1 – 2 – 3 – 4 – 5	Satisfying
Inadequate power	1 – 2 – 3 – 4 – 5	Adequate power
Dull	1 – 2 – 3 – 4 – 5	Stimulating
Rigid	1 – 2 – 3 – 4 – 5	Flexible

Overall reaction to the Touchpad Input:

Terrible	1 – 2 – 3 – 4 – 5	Wonderful
Difficult	1 – 2 – 3 – 4 – 5	Easy
Frustrating	1 – 2 – 3 – 4 – 5	Satisfying
Inadequate power	1 – 2 – 3 – 4 – 5	Adequate power
Dull	1 – 2 – 3 – 4 – 5	Stimulating
Rigid	1 – 2 – 3 – 4 – 5	Flexible

References:

[1]: Venkatesh, V., Morris, M., Davis, G., Davis, F., User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, Vol. 27, No. 3, September 2003, pp. 425-478.

[2]: Chin, J.P., Diehl, V.A., Norman, K.L., Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. *Proceedings of the ACM conference on Human Factors in Computing systems (CHI'98)*, May 15-19, 1988, Washington D.C., United States, pp. 213-218.

REFERENCES

American Foundation for the Blind (2001). *Quick facts and figures on blindness and low vision.* Retrieved May 30, 2005, from American Foundation for the Blind Web site: http://www.afb.org/info_document_view.asp?documentid=1374.

Allport, D.A., Antonis, B., and Reynolds, P. (1972). On the division of attention: a disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology*, 24, 255-265.

Arons, B. (1992). *A review of the cocktail party effect.* Retrieved on May 27, 2007, from MIT Media Lab Web site: http://xenia.media.mit.edu/~barons/pdf/arons_AVIOSJ92_cocktail_party_effect.pdf.

Asakawa, C. and Itoh, T. (1998). User interface of a home page reader. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp.149-156).

Asakawa, C. and Takagi, H. (2000). Annotation-based transcoding for nonvisual Web access. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp.172-179).

Asakawa, C., Takagi, H., Ino, S. and Ifukube, T. (2002). Audiotry and tactile interfaces for representing the visual effects on the Web. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 65-72).

Baddeley, A. D. (1986). *Working memory.* New York, New York: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Science,* 4, 417-423.

Baddeley, A. D. and Hitch, G. J. (1974). Working Memory. In G.A. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory* (Vol. 8, pp. 47-89), New York: Academic Press.

Badler, N., Manoochehri, K. and Baraff, D. (1986). Multi-dimensional Input Techniques and Articulated Figure Positioning by Multiple Constraints. *ACM Workshop on Interactive 3D Graphics* (pp. 151-170).

Bernsen, N.O., Dybkjær, L., and Dybkjær, H. (1994). A dedicated task-oriented dialogue theory in support of spoken language dialogue systems design. In *Proceedings of the international conference on spoken language processing* (pp. 875-878).

Billi, R., Castagneri, G., and Danieli, M. (1996). Field trial evaluation of two different information inquiry systems. In *Proceedings of the IEEE third workshop on interactive voice technology for telecommunications applications* (pp. 129-134).

Billinghurst, M. (1998). Put that where? Voice and gesture at the graphics interface. *Computer Graphics*, 32(4), 60-63.

Boian, R., Sharma, A., Merians, A., Burdea, G., Adamovich, S., Recce, M., Tremaine, M. and Poizner, H. (2002). Virtual reality-based post stroke rehabilitation. In *Proceedings of medicine meets virtual reality* (pp. 64-70).

Bolt, R.A. (1980). Put-that there. *Computer Graphics*, 14 (3), 262-270.

Bolt, R.A. (1987). Conversing with computers. In R. Baecher, W. Buxton, (Eds.). *Readings in human-computer interaction: a multidisciplinary approach*, Los Altos, California: Morgan-Kaufmann.

Brennan, S.E. and Hulteen, E. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, 8, 143-151.

Brewster, S.A., Räty, V.P. and Kortekangas, A. (1996). Earcons as a method of providing navigational cues in a menu hierarchy. In *Proceedings of HCI on people and computers* (pp.169-183).

Brewster, S.A. (1998). Using nonspeech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction*, 5(3), 224-259.

Brewster, S., Lumsden, J., Bell, M., Hall, M. and Tasker, S. (2003). Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 473-480).

Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Doubille, B., Prevost, S. and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics*, 28(4), 413-420.

Chen, C. (1997). Structuring and visualising the WWW by generalised similarity analysis. In *Proceedings of the ACM conference on hypertext and hypermedia* (pp. 177-186).

Chen, X., Chung, J., Lacsina, P. and Tremaine, M. M. (2004). Mobile browsable information access for the visually impaired. In *Proceedings of the tenth American conference on information systems*.

Chen, X., Lacsina, P. and Tremaine, M. M. (2003). Designing nonvisual bookmarks for mobile PDA users. in *Proceedings of the ninth American conference on informaiton systems.*

Chen, X. and Tremaine, M. (2005). Multimodal user input patterns in a non-visual context. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 206-207).

Chen, X., Tremaine, M., Lutz, R., Chung, J. and Lacsina, P. (2006). AudioBrowser: A mobile browsable information access for the visually impaired, *International Journal of Universal Access in the Information Society*, 5(1), pp. 4-22.

Chi, E.H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 490-497).

Colwell, C., Petrie, H., Kornbrot, D., Hardwick, A. and Furner, S. (1998). Haptic virtual reality for blind computer users. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 92-99).

Chin, J.P., Diehl, V.A. and Norman, K.L. (1998). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 213-218).

Cocchini, G., Logie, R. H., Sala, S. D., MacPherson, S. E., and Baddeley, A. D. (2002). Concurrent performance of two memory tasks: evidence for domain-specific working memory systems. *Memory & cognition*, 30(7), 1086-1095.

Cohen, P. (1992). The role of natural language in a multimodal interface. In *Proceedings of the ACM symposium on user interface software and technology* (pp. 143-149).

Cohen, P.R., Dalrymple, M., Moran, D.B., Pereira, F.C.N., Sullivan, J.W., Gargan, R.A., Schlossberg, J.L. and Tyler S.W. (1989). Synergistic use of direct manipulation and natural language. In *Proceedings of the ACM SIGCHI conference on Human Factors in Computing systems* (pp. 227-234).

Cohen, R.F., Haven, V., Lanzoni, J.A., and Meacham, A. (2006). Using an audio interface to assist users who are visually impaired with steering tasks. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 119-124).

Cohen, R.F., Yu, R. Meacham, A., and Skaff, J. (2005). PLUMB: Displaying graphs, to the blind using an active auditory interface. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 182-183).

Dolphin Group. *Hal - Screen reader with speech and braille support.* Retrieved April 15, 2005, from Dolphin: Bring Access to Life Web site: http://www.dolphinuk.co.uk/products/hal.htm.

Edwards, A.D., McCartney, H., and Fogarolo, F. (2006). Lambda: A multimodal approach to making mathematics accessible to blind Students. *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 48-54).

Edwards, W. K. and Mynatt, E. D. (1994). An architecture for transforming graphical interfaces. in *Proceedings of ACM conference on user interface software and technology* (pp. 39-47).

ETS. *Sample questions for Test of English as a Foreign Language (TOEFL).* Retrieved September 2, 2005 from ETS Web site: http://www.ets.org.

Eysenck, M. W. (2004). *Psychology: An international perspective.* Hove, UK and New York, USA: Psychology Press.

Fleetwood, M.D., Byrne, M.D., Centgraf, P., Dudziak, K., Lin, B. and Mogilev, D. (2002). An analysis of text entry in Palm OS: Graffiti and the virtual keyboard. In *Proceedings of human factors and ergonomics society (HFES) 46th annual meeting* (pp. 617-621).

Fraser N. (1997). Assessment of interactive systems. In D. Gibbon, R. Moore, and R. Winski, (Eds.), *Handbook of standards and resources for spoken language systems* (pp. 564-641), New York, NY: Mouton de Gruyter.

Freedom Scientific. *Braille n Speak.* Retrieved April 15, 2005, from Freedom Scientific Web site: http://www.blazie.co.uk/productsBnS.htm.

Freedom Scientific. *JAWS.* Retrieved April 15, 2005, from Freedom Scientific Web site: http://www.freedomscientific.com/fs_products/software_jaws.asp.

Friedlander, N., Schlueter, K. and Mantei, M. (1998). Bullseye! When Fitt's law doesn't fit. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 257-264).

Goldberg, D. and Richardson C. (1993). Touch-typing with a stylus. In *Proceedings of the ACM SIGCHI conference on Human Factors in computing systems* (pp. 80-87).

Goldstein, M., Book, R., Alsio, G. and Tessa, S. (1999). Non-keyboard QWERTY touch typing: A portable input interface for the mobile user. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 32-39).

Grasso, M.A., Ebert, D.S. and Finin, T.W. (1998). The integrality of speech in multimodal interfaces. *ACM Transactions on Computer-Human Interaction*, 5(4), 303-325.

Green, M. (1986). A survey of three dialogue models. *ACM Transactions on Graphics*, 5(3), 244-275.

Guiard, Y. (1987). Asynmmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *The Journal of Motor Behavior*, 19(4), 486-517.

GW Micro, Inc. *Window Eyes*. Retrieved April 15, 2005, from GW Micro Web site: http://www.gwmicro.com/products.

Halverson, D., Horn, D., Karat, C. and Karat, J. (1999). The beauty of errors: Patterns of error correction in desktop speech systems. In *Proceedings of INTERACT'99* (pp. 133-140).

Harper, S., and Patel, N. (2005). Gist Summaries for visually impaired surfers. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 90-97).

Hauptmann, A.G., and McAvinney, P. (1993). Gestures with speech for graphics manipulation. *International Journal of Man-Machine Studies*, 38, 231-249.

Hinckley, K., Pausch, R., Goble, J. and Kasselll, N. (1994). A survey of design issues in spatial input. In *Proceedings of the ACM conference on user interface software and technology* (pp. 213-222).

Hinckley, K., Pausch, R., Goble, J. and Kasselll, N. (1994). Passive real-world interface props for neurosurgical visualization. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 452-458).

Holzapfel, H., Nickel, K. and Stiefelhagen, R. (2004). Implementation and evaluation of a constraint-based mulitmodal fusion systems for speech and 3D pointing gestures. In *Proceedings of international conference on multmodal interfaces* (pp. 175-182).

IBM. *Home Page Reader*. Retrieved April 15, 2005, from IBM Web site: http://www-306.ibm.com/able/solution_offerings/hpr.html.

IBM. *ViaVoice*. Retrieved April 15, 2005, from IBM Web site: http://www-306.ibm.com/software/voice/viavoice/.

Ibrahim, A. and Johansson, P. (2002). Multimodal dialogue systems for interactive TV applications. In *Proceedings of the fourth IEEE international conference on multimodal interfaces* (pp. 117-122).

Immersion Corporation. *CyberGloves*. Retrieved April 15, 2005, from Immersion Corporation Web site: http://www.immersion.com/3d/products/cyber_glove.php.

James, W. (1890). *The principles of psychology*. New York: Henry Holt.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Karat, J., Horn, D.B., Halverson, C.A. and Karat, C.M. (2000). Overcoming unusability: Developing efficient strategies in speech recognition systems. In *Extended abstracts of the ACM SIGCHI conference on human factors in computing systems* (pp. 141-142).

Karat, C., Halverson, C., Horn, D. and Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 568-575).

Karshmer, A.I., Gupta, G., Pontelli, E., Miesenberger, K., Ammalai, N., Gopal, D., Batusic, M., Stoger, B., Palmer, B., and Guo, H. F. (2004). UMA: A system for universal mathematics accessibility. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 55-62).

Költringer, T. and Grechenig, T. (2004). Comparing the immediate usability of Graffiti 2 and virtual keyboard. In *Extended abstracts of the ACM SIGCHI conference on human factors in computing systems* (pp. 1175-1178).

Logie, R.H. (1995). *Visuo-spatial working memory*. Hove, UK: Lawrence Erlbaum Associates.

Mack, R.L., Lewis, C.H. and Carroll, J.M. (1983). Learning to use word processors: Problems and prospects. *ACM Transactions on Office Information Systems*, 1(3), 254-271.

MacKenzie, I.S. and Zhang, S. (1997). The immediate usability of Graffiti. In *Proceedings of graphics interface* (pp. 129-137).

Martin, G.L. (1989). The utility of speech input in user-computing interfaces. *International Journal on Man-Machine Studies*, 30, 355-375.

Martin, P., Crabbe, F., Adams, S., Baatz, E. and Yankelovich, N. (1996). SpeechActs: A spoken-language framework. *IEEE Computer*, 29(7), 33-40.

McGee, M.R., Gray, P.D. and Brewster, S.A. (2000). The effective combination of haptic and auditory textural information. In *Proceedings of the first workshop on haptic human-computer interaction* (pp 33-38).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, Illinois: University of Chicago Press.

McNeill, D. (Ed.) (2000). *Language and gesture*. Cambridge, Massachusetts: Cambridge University Press.

McNeill Lab. *Gesture basics*. Retrieved April 15, 2005, from McNeill Lab, Center for Gesture and Speech Research, University of Chicago Web site: http://mcneilllab.uchicago.edu/topics/basics.html.

McTear, M.F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, 34(1), 90-169.

Merians, A. S., Jack, D., Boian, R., Tremaine, M. M., Burdea, G. C., Adamovich, S., Recce, M. and Poizner, H. (2002). Virtual reality-augmented rehabilitation for patients following stroke. *Physical Therapy*, 82(9), 898-915.

Montgomery, D.C. (2004), *Design and analysis of experiments* (fifth edition). New York, New York: John Wiley & Sons.

Morley, S., Petrie, H., O'Neill, A.M., and McNally, P. (1998). Auditory navigation in hyperspace: Design and evaluation of a non-visual hypermedia system for blind users. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 100-107).

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Education Psychology*, 87(2), 319-334.

Mynatt, E.D. (1994). Auditory presentation of graphical user interfaces. In Kramer, G. (ed), *Auditory display: Sonification, audification and auditory interfaces* (pp. 533-555), Reading, Massachusetts: Addison-Wesley.

Mynatt, E. D. and Edwards, W. K. (1992). Mapping GUIs to auditory interfaces. In *Proceedings of the fifth annual symposium on user interface software and technology* (pp. 61-70).

Mynatt, E. D. and Edwards, W. K. (1995). New metaphors for nonvisual interfaces. In A. D. N. Edwards (Ed.), *Extraordinary human-computer interaction: Interfaces for users with disabilities*. Cambridge, UK: Cambridge University Press.

Navon, D. and Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, 86(3), 214-255.

Nielsen, J. (1994). *Usability Engineering*, San Francisco, California: Morgan Kaufmann.

Nielsen, J., Mack, R.L, Bergendorff, K.H. and Grischkowsky, N.L. (1986). Integrated software in the professional work environment: Evidence from questionnaires and interviews. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 162-167).

Norman, D. (1990). *The design of everyday things*. New York, New York: Doubleday.

Oogane, T. and Asakawa, C. (1998). An interactive method for accessing tables in HTML. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 126-128).

Ostby, E. (1986). Describing free-form 3D surfaces for animation. In *Proceedings of ACM workshop on interactive 3D graphics* (pp. 251-258).

Oviatt, S.L. (1996). Multimodal interfaces for dynamic interactive maps. In *Proceedings of the ACM SIGCHI conference on human factors in computing Systems* (pp. 95-102).

Oviatt, S.L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12, 93-129.

Oviatt, S.L. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42(11), 74-81.

Oviatt, S.L. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the ACM SIGCHI conference on human factors in computing* (pp. 576-583).

Oviatt, S.L. and Adams, B. (2000). Designing and evaluating conversational interfaces with animated characters. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, (eds.), *Embodied conversational agents* (pp. 319-343), Cambridge, Massachusetts: MIT Press.

Oviatt, S.L. and Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3), 45-53.

Oviatt, S.L., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M. and Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the fifth international conference on multimodal interfaces* (pp. 44-51).

Oviatt, S.L., DeAngeli, A. and Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on human factors in computing systems* (pp. 415-422).

Oviatt, S.L., Laniran, Y., Bernard, J. and Levow, G.A. (1998). Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech*, 41(3), 419-442.

Oviatt, S.L. and Olsen, E. (1994). Integration themes in multimodal human-computer interaction. In *Proceedings of the international conference on spoken language processing* (pp. 551-554).

Oviatt, S. and VanGent, R. (1996). Error resolution during multimodal human-computer interaction. In *Proceedings of the fourth international conference on spoken language processing* (pp. 204-207).

Parente, P. (2004). Audio enriched links: Web page previews for blind users. *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp.2-7).

Petrie, H., Morley, S. and Weber, G. (1995). Tactile-based direct manipulation in GUIs for blind users. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 428-429).

Petrie, H., Morley, S., McNally, P., Graziani, P. and Emiliani, P.L. (1996). Access to hypermedia systems for blind students. In D. Burger (Ed.), *New technologies in the education of the visually handicapped.* London, UK: John Libbey.

Petrie, H., Morley, S., McNally, P., O'Neill, A.M. and Majoe, D. (1997). Initial design and evaluation of an interface to hypermedia systems for blind users. In *Proceedings of the eighth ACM conference on Hypertext* (pp. 48-56).

Potjer, J., Russel, A., Boves, L., and Os. E.D. (1996). Subjective and objective evaluation of two types of dialogues in a call assistance service. In *Proceedings of IEEE interactive voice technology for telecommunications applications* (pp. 121-124).

Quek, F., McNeill, D., Bryll, R., Duncan S., Ma, X.F., Kirbas, C., McCullough, K.E. and Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3), 171-193.

Quinn, J. G., and McConnell, J. (1996). Irrelevant pictures in visual working memory. *Quarterly Journal of Experimental Psychology*, 49A(1), 200–215.

Ramloll, R., Yu, W., Riedel, B. and Brewster, S.A. (2001). Using non-speech sounds to improve access to 2D tabular numerical information for visually impaired users. In *Proceedings of the British HCI group annual conference on human-computer interaction* (pp. 515-530).

Robbe-Reiter, S., Carbonell, N. and Dauchy, P. (2000). Expression constraints in multimodal human-computer interaction. In *Proceedings of the fifth international conference on intelligent user interfaces* (pp. 225-228).

Robbins, T. W., Weinberger, D., Taylor, J. G. and Morris, R. G. (1996). Dissociating executive functions of the prefrontal cortex. *Philosophical Transactions: Biological Sciences*, 351(1346), 1463-1471.

Roth, P., Petrucci, L., Assimacopoulos, A., Pun, T. (1998). AB-Web: Active audio browser for visually impaired and blind users. In *Proceedings of international community for auditory display.*

Roth, P., Petrucci, L., Assimacopoulos, A. and Pun, T. (2000). Audio-haptic Internet browser and associated tools for blind and visually impaired computer users. In Proceeding of *workshop on friendly exchanging through the net* (pp. 57-62).

Sanchez, J. and Baloian, N. (2005). Modeling audio-based virtual environments for children with visual disabilities. In *Proceedings of the world conference on educational multimedia hypermedia and telecommunications* (pp. 1652-1659).

Sanchez, J. and Saenz, M. (2005). 3D sound interactive environments for problem solving. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 173-179).

ScanSoft. *Dragon Naturally Speaking*. Retrieved April 15, 2005, from ScanSoft Web site: http://www.scansoft.com/naturallyspeaking/.

Schneider, W. and Shiffrin, R.M. (1977). Controlled and automatic human information processing I: Detection, search, and attention. *Psychological Review*, 84, 1 - 66.

Sears, A., Feng, J. and Oseitutu, K. (2003). Hands-free, speech-based navigation during dictation: Difficulties, consequences, and solutions. *Human-Computer Interaction*, 18, 229-257.

Shaffer, L.H. (1975). Multiple attention in continuous verbal tasks. In S. Dornic (Ed.), *Attention and performance V*, New York, NY: Academic Press.

Shah, P. and Miyake, A. (1996). The separability of working memory resources for spatial thinking and launguage processing: An individual differences approach. *Journal of Experimental Psychology: General*, 125(1), 4-27.

Shiffrin, R.M. and Schneider, W. (1977). Controlled and automatic human information processing II: Perceptual learning, automatic attending, and a general theory, *Psychological Review*, 84, 127-190.

Simon, J. L. (1997). *Resampling: the new statistics* (second edition). Arlington VA: Resampling Stats.

Simon, S. *Parametric versus nonparametric tests*. Retrieved April 15, 2005, from Children's Mercy Hospitals & Clinics Web site: http://www.cmh.edu/ stats/ ask/parametric.asp.

Smith, A.C., Francioni, J.M., Anwar, M., Cook, J.S., Hossain, A., and Rahman, M. (2004). Nonvisual tool for navigating hierarchical structures. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 133-139).

Sohlberg, M.M. and Mateer, C.A. (1989). *Introduction to cognitive rehabilitation: theory and practice*. New York: Guilford Press.

Sowa, T. and Wachsmuth, I. (1999). Understanding coverbal dimensional gestures in a virtual design environment. In *Proceedings of the ESCA workshop on interactive dialogue in multi-modal systems* (pp. 117-120).

Sowa, T. and Wachsmuth, I. (2000). Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. In *Post-proceedings of the conference of gestures: Meaning and use* (pp. 365-376).

Spence, C. and Driver, J. (1997). Audiovisual links in attention: Implications for interface design. *Engineering psychology and cognitive ergonomics*, 2, 185-192.

Sweller, J., Chandler, P., Tierney P., and Cooper, M. (1990). Cognitive load as factor in the structuring of technical material. *Journal of experimental psychology*, 119(2), 176-192.

Suhm, B., Myers, B. and Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction*, 8(1), 60-98.

Takagi, H., Asakawa, C., Fukuda, K. and Maeda, J. (2002). Site-wide annotation: reconstructing existing pages to be accessible. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 81-88).

Thatcher, J. (1994). Screen Reader/2: Access to OS/2 and the graphical user interface. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 39-47).

Thorisson, K., Koons, D. and Bolt, R. (1992). Multi-modal natural dialogue. In *Video proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 653).

Treisman, A. and Davies, A. (1973). Divided attention to ear and eye. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 101-117), New York, New York: Erlbaum.

United States National Highway Traffic Safety Administration. (1997). *An investigation of the safety implications of wireless communication in vehicles*. Retrieved on May 27, 2007, from the Web site of National Highway Traffic Safety Administration: http://www.nhtsa.dot.gov/people/injury/research/wireless/.

Venkatesh, V., Morris, M., Davis, G. and Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478.

Wall, S. and Brewster, S. (2006). Feeling what you hear: Tactile feedback for navigation of audio graphs. In *Proceedings of ACM SIGCHI conference on human factors in computing* (pp. 1123-1132).

Welford, A.T. (1952). The psychological refractory period and the timing of high-speed performance – a review and a theory. *British Journal of Psychology*, 43: 2-19.

Wickens, C. (1980). The structure of attentional resources. In R.S. Nickerson (ed.), *Attention and performance VIII* (pp. 239-257), Hillsdale, New Jersey: Lawrence Erlbaum.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman and R. Davies (eds.), *Varieties of attention* (pp. 63-102), New York, New York: Academic Press.

Wickens, C. (1992). *Engineering psychology and human performance* (Second Edition), New York, New York: HarperCollins.

Wickens, C.D., Gordon, S.E., and Liu, Y. (1998). *An introduction to human factors engineering*. New York, New York: Addison Wesley Longman.

Wickens, C.D. and Liu, Y. (1988). Code and modalities in multiple resources: A success and a qualification. *Human Factors*, 30, 5, 599-616.

Wickens, C. D. and Ververs, P. M. (1998). Allocation of attention with head-up displays. Technical report of Aviation Research Lab, Institute of Aviation, DOT/FAA/AM-98/28.

Wikipedia (2006). *Bootstrapping*. Retrieved September 2, 2006, from Wikipedia Web site: http://en.wikipedia.org/wiki/Bootstrap_%28statistics%29.

Williams, C. and Tremaine, M. (2001). SoundNews: An audio browsing tool for the blind. In *Proceedings of the international conference on universal access in human-computer interaction* (pp. 1029-1033).

Wobbrock, J.O., Aung, H.H., Myers, B.A. and LoPresti, E.F. (2005) Integrated text entry from power wheelchairs. *Journal of Behaviour and Information Technology*, 24 (3), 187-203.

Wobbrock, J.O., Chau, D.H. and Myers, B.A. (2007). An alternative to push, press, and tap-tap-tap: Gesturing on an isometric joystick for mobile phone text entry. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 667-676).

Wobbrock, J.O. and Myers, B.A. (2006). From letters to words: Efficient stroke-based word completion for trackball text entry. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 2-9).

Wobbrock, J.O., Myers, B.A. and Rothrock, B. (2006). Few-key text entry revisited: Mnemonic gestures on four keys. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 489-492).

Wobbrock, J.O., Myers, B.A., Aung, H.H. and LoPresti, E.F. (2004). Text entry from power wheelchairs: EdgeWrite for Joysticks and touchpads. In *Proceedings of the*

international ACM SIGACCESS conference on computers and accessibility (pp. 110-117).

Wobbrock, J.O., Myers, B.A. and Kembel. J.A. (2003). A stylus-based text entry method designed for high accuracy and stability of motion. In *Proceedings of the ACM symposium on user interface software and technology* (pp. 61-70).

Woodhead, M. M. and Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. *Memory and Cognitive*, 9(4), 368-370.

Xiao, B., Girand, C. and Oviatt, S.L. (2002). Multmodal integration patterns in children, In *Preceedings of international conference on spoken language processing* (pp. 629-632).

Xiao, B., Lunsford, R., Coulston, R., Wesson, M. and Oviatt, S.L. (2003). Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In *Proceedings of the fifth international conference on multimodal interfaces* (pp. 265-272).

Xu, S., Fang, X., Brzezinski, J., and Chan, S. (2005). A dual-modal presentation of network relationships in texts. In *Proceedings of the eleventh Americas conference on information systems* (pp. 2348-2356).

Yankelovich, N. (1996). How do users know what to say? *ACM Interactions*, 3(6), 32-43.

Yankelovich, N. (1998). Using natural dialogs as the basis for speech interface design. In S. Luperfoy, (Ed.), *Automated Spoken Dialog Systems*, Cambridge, Massachusetts: MIT Press.

Yankelovich, N., Levow, G.A. and Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces. In *Proceedings of ACM SIGCHI conference on human factors in computing systems* (pp. 369-376).

Yankelovich, N. and Lai, J. (1998). Designing speech user interfaces. In *Extended abstracts of ACM SIGCHI conference on human factors in computing systems* (pp. 131-132).

Yesilada, Y., Stevens, R., Goble, C. and Hussein, S. (2004). Rendering tables in audio: the interaction of structure and reading styles. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 16-23).

Yu, C-H. *Parametric Tests*. Retrieved April 15, 2005, from Arizona State University Web site: http://seamonkey.ed.asu.edu/~alex/teaching/WBI/ parametric_test.html.

Zhao, S., Dragicevic, P., Chignell, M., Balakrishan, R. and Baudisch, P. (2007). EarPod: Eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 1395-1404).

Zhao, H., Plaisant, C., Shneiderman, B. and Duraiswami, R. (2004). Sonification of geo-referenced data for auditory information seeking: design principle and pilot study. In *Proceedings of the international conference on auditory display*.

Zhao, H., Plaisant, C. and Shneiderman, B. (2005). iSonic: Interactive sonification for non-visual data exploration. In *Proceedings of the international ACM SIGACCESS conference on computers and accessibility* (pp. 194-195).